

CARNEGIE MELLON UNIVERSITY

Value and Trade-offs in Learning from Consumer Location Data

Author:
Meghanath M Y

Committee:
Prof. Beibei LI (Chair)
Prof. Anindya GHOSE
Prof. Rema PADMAN

*A thesis submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy*

in

Information Systems and Management
Heinz College
Carnegie Mellon University



April 27, 2021

Value and Trade-offs in Learning from Consumer Location Data

Meghanath M Y

*A thesis submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in*

Information Systems and Management

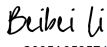
Heinz College

Carnegie Mellon University

Thesis Defended On:

Tuesday, April 27, 2021.

Approved by the Committee:

DocuSigned by:

C905A053574643A...

4/29/2021

Prof. Beibei Li,
Carnegie Mellon University

Date


DocuSigned by:

A8B20EC0B077480...

4/29/2021

Prof. Anindya Ghose,
New York University

Date

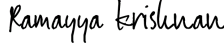
DocuSigned by:

41A4E11F82C3438...

4/29/2021

Prof. Rema Padman,
Carnegie Mellon University

Date

Approved by the Dean:

DocuSigned by:

7D2BBEF5A516439...

4/29/2021

Prof. Ramayya Krishnan,
Dean, Heinz College
Carnegie Mellon University

Date

Declaration of Authorship

I, Meghanath M Y, declare that this thesis titled, "Value and Trade-offs in Learning from Consumer Location Data" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: *Meghanath*

Date: April 27, 2021

To my parents, Macha Geetha and Macha Yadagiri

Abstract

Value and Trade-offs in Learning from Consumer Location Data

by Meghanath M Y

Location data has changed the way we understand human behavior. 76% of the population in the advanced economies own a smartphone. These percentages continue to rocket. The fast penetration of smartphones, combined with the wide adoption of location services, has produced a vast volume of behavior-rich mobile consumer location data. This thesis presents three case studies in which we propose novel methods to discern the value and trade-offs in learning consumer behavior from location data.

In Chapter 1, we study the privacy-utility trade-off associated with the collection and sharing of consumer location data. We find that high privacy risks prevail in the absence of obfuscation on the shared location data. We propose a novel framework enabling a data collector to balance the privacy-utility trade-off. We empirically demonstrate the performance of this approach on smartphone location data of 40,000 consumers collected across several weeks.

In Chapter 2, we study the explainability and predictive accuracy trade-off in learning from location data. We present x-PACS, a new sub-space search learning algorithm that jointly explains and detects anomalous patterns. Explanations are useful in making learning algorithms more transparent to the data practitioner. Using several real-world data sets, we show the effectiveness of x-PACS in anomaly explanation over various baselines and demonstrate its competitive predictive performance.

In Chapter 3, we investigate the value proposition of discerning social determinants of health from location data. We identify individual lifestyles, accessibility to health care facilities, neighborhood characteristics, and socioeconomic factors from location data. We successfully uncover heterogeneous lifestyles of over 10,000 (anonymous) Baltimore and D.C. residents. Our framework reveals that an individual's lifestyle choice is a critical predictor of future hospitalization. Importantly, *regularity*, rather than total time spent at healthy and unhealthy activities, predict future hospitalization. Comparison of the proposed learner that jointly represents several social determinants with various baselines shows its superiority.

Acknowledgements

This thesis would not be possible without the support and guidance I received from many people over the past five years.

I would first like to thank my advisor Beibei Li for her continuous guidance in helping me formulate research questions and methodology. I always felt like I was receiving feedback from a friend during our discussions. I am grateful to her for teaching me to be open to new streams of research and sharpen my thinking.

I would like to thank Anindya Ghose for inspiring me to be a better researcher every time I chatted with him. Thanks to Rema Padman, who has been crucial in helping me structure the last chapter of my thesis.

A huge thank you to my co-authors. Natasha Zhang Foutz, who got me access to location data, without which this thesis would not be possible. Leman Akoglu, for teaching me the importance of presentation in a research paper.

I would like to thank Alessandro Acquisti, Lee Branstetter, Jon Caulkins, George Chen, David Choi, Alexandra Chouldechova, Christos Faloutsos, Pedro Ferreira, Martin Gaynor, Brian Kovak, Ramayya Krishnan, Michael Smith, and Rahul Telang, for their advice and encouragement throughout the years. A special thanks to Michelle Wirtz and Emily Marshall for making this journey easier.

To my long-time mentors, Deepak Pai, Ritwik Sinha who inspired me to pursue a Ph.D. To my undergraduate friends, Aravind Atreya, Manoj Babu, Animesh Kumar, Shouvik Mukherjee, Siva Praneeth, and Nithin Reddy, who let me vent without complaining too much. To my Pittsburgh friends, Nathaniel Breg, Yangfan Liang, Emaad Manzoor, Siddhartha Sharma, Shubranshu Shekhar, and Shruthi Venkatesh, for the thoughtful discussions and for indulging me in fun outside work.

A special thank you to my fiancée, Danielle Ourada, for her love, care, and constant support. Finally, I would like to extend my gratitude to my siblings : Vani, Veena; my Mom, Geetha; Dad, Yadagiri for teaching me by example to work hard and aim higher.

Contents

Declaration of Authorship	iii
Abstract	vii
Acknowledgements	ix
1 Personalized and Interpretable Privacy Preservation	1
1.1 Introduction	1
1.1.1 Smart Tracking, Targeting, and Privacy	1
1.1.2 Research Agenda and Challenges	3
1.1.3 Overview of Proposed Framework	4
1.1.4 Summary of Key Findings	6
1.1.5 Summary of Key Contributions	7
1.2 Literature Review	8
1.2.1 Literature on Consumer Privacy	8
1.2.2 Privacy-preserving Methodology I: Syntactic Models	10
1.2.3 Privacy-preserving Methodology II: Differentially Private Algorithms	11
1.2.4 Location-based Mobile Marketing	13
1.3 Data	13
1.4 Methodology	15
1.4.1 Quantification of Consumer's Privacy Risk	17
Trajectory Feature Extraction.	17
Sensitive Attribute Inference.	20
Re-identification Threat.	20
1.4.2 Quantification of Advertiser's Utility	21
1.4.3 Obfuscation Scheme	23
1.5 Empirical Study	25
1.5.1 Quantification of Consumer's Privacy Risk	25
1.5.2 Quantification of Advertiser's Utility	27
1.5.3 Obfuscation Scheme for Data Collector	28
1.5.4 Model Comparison	28
Comparison to Rule-based Obfuscations.	29
Comparison to Risk-based Obfuscations.	30
Comparison to Latest Suppression Models.	31
1.6 Robustness Tests	32
1.6.1 Alternate Utility Function : Activity Prediction	32
Architecture	33
Training, Model Selection and Hyperparameter tuning	33
Utility Measurement	34
1.6.2 Suppression based on Recency and Time Spent	34

1.6.3	Varying Sample Sizes	35
1.6.4	Varying Dimensionality	35
1.7	Conclusion	36
2	Explaining Anomalies in Groups	39
2.1	Introduction	39
2.1.1	Desiderata for Anomaly Description	40
2.1.2	Limitations of Existing Techniques	41
2.1.3	Summary of Contributions	41
2.2	Overview and Problem Statement	42
2.2.1	Example x -PACS input-output	42
2.2.2	Main Steps	43
2.2.3	Notation and Definitions	44
2.2.4	Problem Statement	44
2.3	x -PACS: Explaining Anomalies in Groups	45
2.3.1	Subspace Clustering: Finding Hyper-rectangles	45
2.3.2	Refining Hyper-rectangles into Hyper-ellipsoids	48
2.3.3	Summarization: Pack Selection for Shortest Description MDL formulation for encoding a given <i>packing</i>	50
	MDL objective function	52
	Subset selection algorithm	52
2.3.4	Overall Algorithm x -PACS	54
2.4	Experiments	55
2.4.1	Effectiveness of Explanations	55
	Case Studies	56
	x -PACS vs. Rule Learners	59
	Ablation Study	61
2.4.2	Detection Performance	62
2.4.3	Scalability	64
2.5	Related Work	64
2.6	Conclusion	67
3	Social Determinants of Health	71
3.1	Introduction	71
3.1.1	Social Determinants of Health	71
3.1.2	Location Data	71
3.1.3	Research Gap	72
3.1.4	Overview of Proposed Methodology	73
3.1.5	Key Findings	73
3.2	Related Work	74
3.2.1	Behavioral Routine and Activities: Smartphone Data	74
	Surveys and Health Records	75
3.2.2	Behavior as Health Determinants:	75
3.3	Framework	76
3.3.1	Locations to Activity Trajectories	77
3.3.2	Lifestyle Identification	78
	Author Topic Model:	78

Activity trajectories to Lifestyles:	79
3.3.3 Other Social Determinants	80
3.3.4 Health Risk Quantification	81
Modelling Hospitalization	81
Proposed Learner	82
3.4 Data	85
3.5 Empirical Study	87
3.5.1 Lifestyles	87
3.5.2 Health Risk Quantification	89
3.6 Conclusion	93
A Personalized and Interpretable Privacy Preservation	95
A.1 Objective Function Analysis	95
A.2 Early Stopping	96
A.3 Complexity Analysis	98
A.4 Speed-up Heuristic.	99
A.5 Utility Measurement	100
A.6 Model Choices in the Proposed Framework	100
A.7 Additional Literature Review	102
A.8 Sensitive Attribute : Consumer Operating System	103
B Social Determinants of Health	107
B.1 Location to Activity trajectories	107
B.2 Author Topic Models Primer:	108
B.3 D.C. Residents	108
B.4 Robustness Checks	111
Bibliography	115

List of Figures

1.1	Overview of the proposed framework	5
1.2	An example of a consumer’s footprints with 732 unique locations over the five-week sample period	14
1.3	Personalized Risk Management Insights	26
1.4	Proposed framework - $MAP@k$ and $MAR@k$ for varying p	27
1.5	Proposed framework vs risk-based obfuscations - $MAP@1$ and $MAR@1$	30
1.6	Proposed framework : Alternate Data Utility Function	34
1.7	Proposed framework : Home address inference, suppression by recency and time spent.	35
1.8	Proposed framework : Home address inference, varying sample sizes	36
1.9	Proposed framework : Home address inference, varying sample sizes	36
2.1	Example x -PACS input–output.(best in color)	43
2.2	Identifying candidate hyper-rectangles in 1-d (equivalent to intervals) by KDE for varying quantile thresholds q	47
2.3	Example illustration of refining hyper-rectangles to ellipsoids in 2-d. Anomalous points (black) captured by SUBCLUS (Alg. 1) in a (green) rectangle, other anomalous points (blue) in the vicinity, and normal points (red).	48
2.4	x -PACS’s description cost of anomalies in image datasets for $K = 1, \dots, 5$. Naïve/base cost ($K = 0$) is shown w/ a horizontal line per dataset. x -PACS finds the appropriate number of patterns automatically and significantly reduces the description cost.	57
2.5	x -PACS achieves the best balance between interpretability measures (a)–(d) [lower is better for all of them], and is significantly better at detection (e) [higher is better], as compared to several rule learners.	60
2.6	x -PACS scales linearly with input size.	65
3.1	Probabilistic Graphical model of Author Topic Model using plate notation.	78
3.2	Architecture diagram of the proposed learner.	82
3.3	Illustrations of CLSTM and Concatenate layers.	84
3.4	Row-normalized heatmap of activity occurrences (Baltimore). Darker reds indicate higher occurrences.	86
3.5	Weekday Lifestyles (Baltimore)	87
3.6	Weekend Lifestyles (Baltimore)	88

3.7 (Baltimore) Association with Hospitalization : Model free analysis (<i>hospitalization</i>)	89
A.1 Home address inference, POI prediction	99
A.2 Proposed framework model choices	101
A.3 Proposed framework - $MAP@k$ and $MAR@k$ for varying p , OS inference	103
A.4 Proposed framework vs risk-based obfuscations - $MAP@1$ and $MAR@1$, OS inference	104
B.1 DC Weekday Lifestyles	109
B.2 DC Weekend Lifestyles	110
B.3 (D.C.) Association with Hospitalization : Model free analysis (<i>hospitalization</i>)	111

List of Tables

1.1	Comparison of Proposed Method to Relevant Literature in Privacy Preserving Location Data Sharing	13
1.2	Summary statistics of the location data sample under analysis	14
1.3	Description of consumer mobility features	19
1.4	Alternative Schemes: Rule-based Obfuscation	29
1.5	LSUP and GSUP comparison. (Green/Red indicate proposed framework provides a better/worse trade-off)	32
1.6	Activity groups	33
2.1	Dataset statistics. x-PACS achieves significant savings (in bits) by explaining anomalies in groups.	56
2.2	DigitI ‘0’ vs. ‘7’: x-PACS finds one 2-d <i>pack</i>	58
2.3	DigitII ‘8’ vs. ‘2’,‘3’: x-PACS’s one 4-d <i>pack</i>	58
2.4	BrCancer: x-PACS finds five 1-d or 2-d <i>packs</i>	59
2.5	Interpretability measures (a)–(d): x-PACS vs. Rule learners. Also given for reference is detection performance in AUPRC (See §2.4.2 for details).	59
2.6	Rule learners and DT (with respective depths 1–5) compared to x-PACS across datasets on interpretability measures (a)–(d) [all lower the better] as well as detection performance AUPRC [higher the better]. RuleFit leads to underspecified regression in Arrythmia and Yeast which we denote by NA.	62
2.7	Ablation Study: x-PACS vs. ablated x-PACS (no refinement to ellipsoids). Coverage of anomalous points (higher is better), coverage of normal points (lower is better), and % savings (higher is better).	63
2.8	Area under precision-recall curve (AUPRC) on anomaly detection.	64
2.9	Comparison of related work in terms of properties D1–D5 in reference to our Desiderata (see §2.1.1).	69
3.1	Activity groups	77
3.2	Definition and Summary Statistics of Social Determinants and Health Events	83
3.3	Summary statistics of the activity trajectories	86
3.4	(Baltimore) Hospitalization Logit Analysis	90
3.5	(Baltimore) Hospitalization : Additional Logit Analysis	91
3.6	(Baltimore) Hospitalization prediction	91
3.7	(D.C.) Hospitalization Prediction	92
A.1	Early stopping heuristic : POI@k	98
A.2	Early stopping heuristic : Activity Prediction@k	98

A.3	Clock time of the proposed heuristic	99
A.4	Alternative Schemes: Rule-based Obfuscation (Operating System Inference)	104
A.5	LSUP and GSUP comparison : OS inference (Green/Red indicate proposed framework provides a better/worse trade-off)	105
B.1	(D.C.) Hospitalization Logit Analysis	112
B.2	(D.C.) Hospitalization : Additional Logit Analysis	112
B.3	(Baltimore) Logit Analysis : Robustness check for Hospitalization	113
B.4	(D.C.) Logit Analysis : Robustness check for Hospitalization	113

Chapter 1

Personalized and Interpretable Privacy Preservation

1.1 Introduction

1.1.1 Smart Tracking, Targeting, and Privacy

According to the latest Pew Research (Taylor and Silver, 2019), 76% and 45% of the current population in the advanced and emerging economies, respectively, own a smartphone. These percentages continue to rise rapidly. Among the U.S. smartphone consumers, over 90% use location services such as Google Maps (Pew, 2016). The fast penetration of smartphones, combined with the wide adoption of location services, has produced a vast volume of behavior-rich mobile location data (or location data, trajectory data hereafter). These data represent one of the latest, and most important, information sources available to marketers in the evolution of marketing data, from surveys to online clickstream and social media data (Wedel and Kannan, 2016). It has also opened up \$21 billion sales opportunities for advertisers, ranging from e-commerce retailers sending discount coupons to individuals in the vicinity of a store, commonly known as geo-fencing, to personal injury lawyers targeting those in emergency rooms (Luo et al., 2014; Andrews et al., 2016; Ghose, Li, and Liu, 2018; Kelsey, 2018).

Geo-marketing based on mobile location data is attractive to advertisers for multiple reasons. First, mobile location data are straightforward to collect, an app permission away, tracked in the background in most mobile ecosystems, and readily accessible to advertisers.¹ Second, mobile location data are superior to alternative location data. The built-in sensors of mobile devices can provide continuous tracking of the movement trajectory of an individual (i.e., a sequence of fine-grained GPS coordinates). Such individual-level trajectory data are more precise and granular than social media geo-tags and consumer self check-ins. They are also more representative of the population than the less granular taxi and public transportation location data. Third, mobile location data offer an extensive profile of a consumer and portray rich contexts of a consumer's behavior and brand preference, broad lifestyle, socioeconomic status, and social relationship (Ghose, Li, and Liu, 2018). Such offline data become even more powerful if combined with a consumer's online footprints, such as click stream data or social media data,

¹While both Apple and Android have taken measures to limit the collection of location data, guidelines remain ambiguous about the sales of such data to advertisers (Apple, 2014; Verge, 2019).

rendering a holistic online-offline consumer profile. Fourth, excellent location tracking and targeting across apps simplifies ad attribution of a location-based ad campaign. Each advertiser has access to a unique device ID associated with each smartphone, thus benefiting from reduced overhead to stitch together a consumer's location data across sessions or apps and enjoying a holistic view of each consumer when measuring a campaign's effectiveness (Apple, 2012). Fifth, geo-marketing by a butler advertiser also benefits consumers (Ghose, 2017), such as allowing consumers to receive enhanced services, personalization (Chellappa and Shivendu, 2010), and financial benefits such as coupons (Luo et al., 2014; Ghose, Li, and Liu, 2018) or lower insurance premiums (Soleymanian, Weinberg, and Zhu, 2019).

Mobile location data not only provide utility to an advertiser whose butler actions further benefit consumers, but also monetization opportunities to a location data collector who shares the data with the advertiser. Despite of the existence of diverse sources and varieties of mobile location data, the backbone of this rapidly growing mobile location data economy is the huge number of mobile apps. App owners and location data aggregators serve a two-sided market with consumers on one side and advertisers on the other, collecting location data to offer better services to consumers and to monetize with advertisers. For example, a recent article by the New York Times reported that mobile location data collectors accrue half to two cents per consumer per month from advertisers (Valentino-Devries et al., 2018).

Meanwhile this powerful new form of human movement data offers important utility to an advertiser, and thus benefits to consumers and the data collector as well, they entail major privacy risks, such as home location inference. "Privacy" is defined as "the quality or state of being apart from company or observation" in Merriam-Webster. In business contexts, privacy broadly pertains to the protection of individually identifiable information online or offline, and the adoption and implementation of privacy policies and regulations. It is a key driver of online trust (Hoffman, Novak, and Peralta, 1999). More than three-quarters of consumers believe that online advertisers hold more information about them than they are comfortable with; and approximately half of them believe that websites ignore privacy laws (Dupre 2015). For offline location data, privacy risks are exemplified by identifications of a consumer's home address, daily trajectories, and broad lifestyle, as vividly depicted by two recent New York Times' articles (Valentino-Devries et al., 2018; Thompson and Warzel, 2019). These risks are arguably more concerning than those associated with other forms of consumer data, such as an individual's media habit or social media content.

The discussion so far calls for any data collector, before sharing location data with an advertiser, to maintain a crucial trade-off between the utility to the advertiser and privacy risk to a consumer. This responsibility falls primarily upon data collectors as they are situated right between advertisers and consumers, and hold vested interests in continuously maintaining consumers' trust in order to collect and monetize location data.² This notion is also consistent with the extant literature across multiple disciplines on data sharing (Li et al., 2012; Terrovitis, Mamoulis, and Kalnis, 2008; Li

²Cambridge Analytica's misuse of consumer data exemplifies severe backlash on the data collector, Facebook, whose privacy practices resulted in a loss of both consumers and advertisers (Pew, 2018).

and Sarkar, 2009; Chen et al., 2013; Yarovoy et al., 2009; Machanavajjhala, Gehrke, and Götz, 2009). The unique properties of, and hence challenges entailed by, the increasingly accessible and important mobile location data (to be detailed next), nonetheless, call for novel frameworks to accomplish the risk-utility trade-off (or privacy-utility trade-off hereafter). We thus aim to develop a personalized, privacy-preserving framework that incorporates consumer heterogeneity and optimizes a data collector's risk-utility trade-off.

1.1.2 Research Agenda and Challenges

As discussed earlier, there are three key entities in our business setting.

1. *Consumer*: is an individual who owns a smartphone with one or more of the apps installed that transmit the individual's location data to the data collector. Each consumer has the option to opt out of any app's location tracking, with some potential downsides of restricted use of certain app functions, such as maps or local restaurant finders.
2. *Advertiser*: is a firm that acquires data from a data collector to improve the targetability of its marketing campaigns. A subset of advertisers, or even a third party, with access to the location data, might have a stalker intent (stalker hereafter) to perform malicious activities on the location data that invade consumer privacy, such as overly aggressive marketing or ignoring privacy concerns.
3. *Data collector*: is an app owner that collects consumers' location data from its mobile app, or a data aggregator that integrates location data from multiple apps. The data are collected in real time and may then be shared with or sold to advertisers interested in targeting the consumers.

In this work, we take a data collector's perspective and propose a framework for the data collector to balance between protecting consumer privacy and preserving a butler advertiser's utility such as POI recommendation (Muralidhar and Sarathy, 2006). We aim to answer the following essential questions.

1. *Consumer's privacy risk*: What are some of the key privacy risks of mobile location data to a consumer due to an advertiser's potential stalker intent? Can these risks be quantified at a consumer level? Since a data collector has limited purview of how an advertiser could infer a consumer's private information from location data, understanding and quantifying the risks associated with various types of stalker behaviors (or threats hereafter) is a crucial first step.
2. *Advertiser's utility*: What is the value of an obfuscated data set to a butler advertiser's utility? Specifically, what types of key behavioral information can a butler advertiser extract from the data to service or target consumers in a mutually beneficial way?
3. *Data collector's trade-off between consumers' privacy risks and advertiser's utility*: Is there a reasonable privacy-utility trade-off? If yes, what are the necessary steps for the data collector to take?

To accomplish the above, several methodological challenges need to be overcome. From a methodological standpoint, our research questions broadly fall under the paradigm of Privacy-Preserving Data Publishing (PPDP) widely studied in the context of relational databases (Fung et al., 2010). Nonetheless, the unique properties of mobile location data, such as high dimensionality (due to a large number of locations visited), sparsity (few overlaps of locations across consumers), and sequentiality (order of locations visited), pose additional challenges (Chen et al., 2013). For example, traditional k -anonymity, which ensures an individual’s record is indistinguishable from at least $k - 1$ records, and its extensions face the curse of high dimensionality while dealing with granular, sometimes second-by-second location data (Aggarwal, 2005). ϵ -differential privacy anonymization, which ensures adding or deleting a single consumer record has no significant impact on analysis outcomes, and other randomization-based obfuscation techniques (Machanavajjhala et al., 2006), fail to preserve the truthfulness of location data, rendering obfuscated data less useful for an advertiser’s visual data mining tasks. More recent local obfuscation techniques (Chen et al., 2013; Terrovitis et al., 2017) that suppress locations with lower risk-utility trade-off provide a good privacy-utility balance. However, the obfuscation mechanisms are often complex for a data collector to interpret and apply in practice. For instance, the $(K, C)_L$ privacy framework (Chen et al., 2013) requires multiple parameters from a data collector, such as the probability thresholds of a privacy threat to succeed in different types of behaviors. LSUP (Terrovitis et al., 2017) requires similar input parameters. Given the complex nature of these approaches, understanding and setting such parameters are non-trivial for a data collector. Hence, a more interpretable framework is needed to assist a data collector.

Furthermore, the extant approaches do not tie a butler advertiser’s utility to any specific business use case. These approaches, devised mostly from the Computer Science literature, measure an advertiser’s utility with simply the number of unique locations or location sequences preserved in the obfuscated data (Chen et al., 2013; Terrovitis et al., 2017). These measures are rather rudimentary and impractical for an advertiser to interpret or link to monetary decision-making. This challenge thus needs to be tackled by tying the advertiser’s utility to real-world business contexts. We will next overview the proposed framework that intends to address the above challenges.

1.1.3 Overview of Proposed Framework

We provide a brief overview of the proposed framework that consists of three main components: quantification of each consumer’s privacy risk, quantification of an advertiser’s utility, and obfuscation scheme for a data collector.

Quantification of Consumer’s Privacy Risk. While the proposed framework may accommodate a variety of privacy risks, we illustrate the framework by computing two specific risks of vital concerns to consumers. One is “sensitive attribute inference”, where a consumer’s sensitive attributes, such as home address, is being inferred (Li, Shirani-Mehr, and Yang, 2007; Tucker, 2013; Gardete and Bart, 2018; Rafieian and Yoganarasimhan, 2018). And the other is “re-identification threat”, where all locations visited by a consumer

are being identified based on a subset of the locations (Samarati, 2001; Pelungrini et al., 2018).

Quantification of Advertiser’s Utility. While the utility of a mobile location data set to an advertiser is multi-faceted, we demonstrate one specific utility related to one arguably most popular and essential business goal examined by the literature – Point-of-Interest (POI hereafter) recommendation in mobile advertising (Ghose, Li, and Liu, 2018). Reliable predictions of a consumer’s future locations would enable an advertiser to target the consumer with context relevant contents and lead to higher business revenues (Ghose, Li, and Liu, 2018). For instance, if a chain restaurant can accurately predict that a consumer is going to be in the vicinity of one of its outlets, it may target the consumer with a discount coupon of value to the consumer. We hence quantify this utility as the accuracy of a similarity-based collaborative filtering recommendation model trained on the location data. The central idea of this recommender is to identify other consumers with similar historical behaviors in order to infer the focal consumer’s future behavior (Bobadilla et al., 2011). Note that while focusing on POI recommendation, our framework can easily accommodate other utility measurements with various business goals as well.

Obfuscation Scheme for Data Collector. Acknowledging many potential solutions to the privacy-utility trade-off may emerge, we propose an obfuscation scheme grounded on the idea of suppressing a subset of a consumer’s locations, given the consumer’s specific privacy risk and the frequency, recency, and time that the consumer spent at each location. We achieve this by introducing consumer-specific parameters that control the number and identities of the locations suppressed for each consumer. The suppression, while reducing each consumer’s privacy risk, also adversely impacts an advertiser’s utility. Hence, we empirically identify the parameters that balance the privacy-utility trade-off through a structured grid search while leveraging the risk quantification for each consumer.

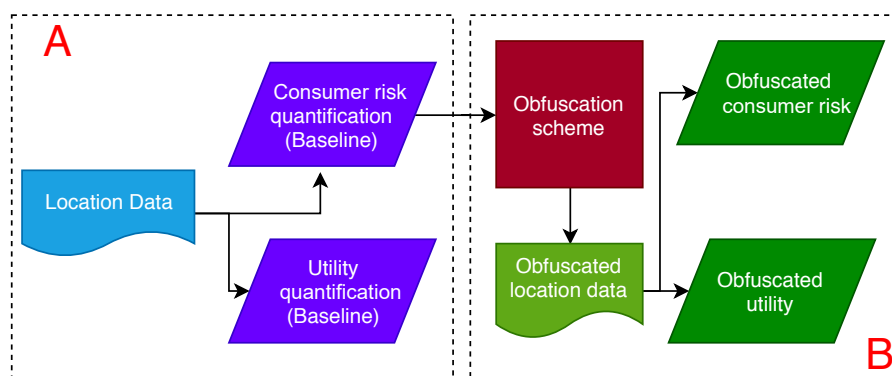


FIGURE 1.1: Overview of the proposed framework

In summary, Figure 1.1 illustrates the proposed framework encompassing the three components discussed above. In Part A, we compute each consumer’s baseline risk and the advertiser’s baseline utility from the original, unobfuscated mobile location data (i.e., the full sample). These would also represent the counterfactual case when no privacy protection is performed. We expect the unobfuscated full sample to yield the maximum utility to the

advertiser, yet incur the maximum privacy risk to the consumers. In Part B, we perform consumer-level obfuscation based on the suppression probabilities of each location computed from each consumer's baseline risk, measures of the informativeness of each location, and a grid parameter. We then calculate the mean risk and utility across all consumers from the obfuscated data. Finally, we repeatedly obfuscate the original data by varying the grid parameter and recalculate the mean risk and utility on the corresponding obfuscated data to empirically determine the best risk-utility trade-off for the data collector. We will describe the details in the Methodology section.

As alluded to earlier, while we illustrate the power and value of the proposed framework by examining two key types of privacy risks and one key advertiser application, the framework is flexible to accommodate other types of privacy risks, such as location sequence or visit frequency inference, for which the risk may be quantified either analytically or via machine learning heuristics (Pellungrini et al., 2018). The framework may accommodate other types of advertiser use cases as well, for which the utility may be computed as the predictive accuracy of the specific business application of interest, such as when a consumer is most likely to convert into a paying customer given prior trajectories, or how much is an advertiser's incremental revenue from geo-marketing. The framework is also applicable to other contexts, for instance, when the data collector conducts geo-marketing for itself or for advertisers without sharing location data. We will summarize the key findings next.

1.1.4 Summary of Key Findings

We validate the proposed framework on a unique data set of nearly one million mobile locations from over 40,000 individuals in a major mid-Atlantic metropolitan area in the U.S. over a period of five weeks in 2018. The main findings are summarized as follows.

First, we find that the absence of an obfuscation scheme, that is, no steps taken by a data collector to ensure consumer privacy, indeed entails high privacy risks to consumers. On average, the success probability is 84% for inferring a consumer's home address and 82% for inferring mobile operating system³. On average, a consumer's home address can be predicted within a radius of 2.5 miles. Moreover, a consumer's entire location trajectories can be fully identified with a 49% success by knowing merely two randomly sampled locations visited by the consumer. It is noteworthy that these success probabilities of various privacy threats are all estimated based on machine learning heuristics, which require only the consumers' locations and corresponding timestamps as the inputs, as we will describe later. Hence, any entities, including advertisers, who have access to the location data could accomplish the same inferences.

Second, we find great value of the mobile location data to an advertiser. An advertiser aiming to target a consumers would be able to predict the next location most likely visited by the consumer with 25% success. This means

³Previous studies have shown a strong relationship between mobile operating system and consumer demographics (eMarketer, 2013).

that by analyzing the behavioral patterns revealed by the historical trajectories, for every one out of four customers, the advertiser is able to design a highly precise geo-targeting strategy.

Finally, a data collector could curtail the potential invasion of consumer privacy by performing data obfuscation. Using the proposed obfuscation scheme, where we suppress each consumer's locations based on the consumer's privacy risks and frequency, recency, and time spent at each location, a data collector may choose from multiple options of risk-utility trade-off via a grid parameter to perform the obfuscation. Moreover, we find that the proposed framework presents a better choice set of risk-utility trade-off when compared to eight baselines obfuscation schemes of various types, including the rule based, consumer risk based, and latest suppression techniques such as Terrovitis et al., 2017. For instance, when the privacy threat is to predict a consumer's home address, the proposed obfuscation scheme reduces the risk by 15%, which represents the maximum decrease when compared to the baselines, with a minimum decrease of less than 1% in an advertiser's utility. We will present a more detailed discussion of the empirical findings and comparisons with the baseline obfuscation schemes in Section 1.5.

1.1.5 Summary of Key Contributions

We propose an interpretable framework built upon the principle of personalized data obfuscation for the emerging and increasingly critical mobile location data. These data exhibit distinctive properties, such as high dimensionality (resulting from massive numbers of locations), sparsity (with few overlaps across visited locations), and sequentiality (with temporal ordering of visited locations), hence imposing unique methodological challenges.

Conceptually, this research demonstrates the importance for any location data collector to preserve both consumer privacy and advertiser utility on a two-sided market. It hence presents a systematic framework to accomplish this privacy-utility balance. It also stands among the first research to demonstrate the immense business values of the novel mobile location data that capture granular human movements and are increasingly leveraged by marketers and other entities, such as municipalities (e.g., for smart city planning). This research simultaneously illustrates the significant privacy risks associated with these data if no framework were in place to preserve consumer privacy.

Managerially, this framework tackles three inter-related critical challenges facing a location data collector: quantification of each consumer's privacy risk, quantification of an advertiser's utility (i.e., value of mobile location data to an advertiser), and design of an intuitive and interpretable obfuscation scheme for a data collector. The framework requires only a single parsimonious input yet offers a data collector multiple, interpretable, and personalized options to protect consumer privacy while preserving an advertiser's utility, hence the data collector's overall monetization opportunity.

Methodologically, this framework (1) quantifies the privacy risk at a consumer level, instead of an aggregate or location level; and quantifies each consumer's privacy risk by extracting a comprehensive set of features from the mobile location data, thus accommodating various types of privacy risks

and allowing identifications of which features contribute the most to the privacy risks; (2) measures an advertiser's utility associated with specific, real-world business use cases, such as POI recommendation shown to improve retailers' incremental revenues (Ghose, Li, and Liu, 2018); (3) proposes an interpretable obfuscation scheme that requires merely one input from the data collector and suppresses locations at each consumer level to furnish the data collector with multiple intuitive options to maintain the privacy-utility trade-off; (4) demonstrates efficacy by validating the proposed framework on a massive, real-world mobile location data set and comparing with eight benchmarks.

Striking a balance between consumer privacy and geo-marketing constitutes part of a broader debate over tracking and targeting on digital platforms. This debate has resulted in actions from both industries and regulatory bodies. For instance, Apple, with 44.6% US smartphone market share (Statista, 2018), introduced limited ad tracking (LAT) in 2016, which allowed consumers to opt out of tracking indefinitely (Apple, 2016). Following suit, Android, the second most adopted mobile ecosystem, rendered more controls to each consumer to limit tracking in its latest software update (Verge, 2019). European Union's General Data Protection Regulation (GDPR Regulation, 2016), effective from May 2018, requires individuals to opt-in (rather than opt out of) behavioral targeting and to give explicit permission for their data to be shared across firms.

Balancing the benefit and privacy risk of consumer location data is increasingly becoming a key concern and top priority for firms and regulatory bodies. Besides strengthening privacy regulations, more research is called for to develop privacy-friendly data storage, processing, and analysis technologies (Wedel and Kannan 2016). Against this background, our research provides empirical evidence and practical solutions to inform the ongoing debate over mobile location tracking and location-based targeting.

The rest of the manuscript is organized as follows. In Section 3.2, we review the literatures from various disciplines that are relevant to our research questions. In Section 3.4, we provide details of our business setting and discuss sampling and summary statistics of the mobile location data under analysis. Section 3.3 describes the details of the proposed framework (Figure 1.1). In Section 1.5, we discuss the empirical results and advantages of the proposed framework. We offer the concluding remarks in Section 1.7.

1.2 Literature Review

We will concisely review the most relevant Marketing, Management, Information Systems (IS), and Computer Science (CS) literature on consumer privacy, privacy-preserving methodologies, and location-based mobile advertising.

1.2.1 Literature on Consumer Privacy

The literature, particularly from Marketing, has a historical, and newly revived, interest in consumer privacy. As different forms of consumer data emerge over time, the literature has examined consumer privacy concerns

that arise from many business contexts and data forms, such as marketing research like surveys (Mayer and White Jr, 1969; De Jong, Pieters, and Fox, 2010; Acquisti, John, and Loewenstein, 2012), direct marketing via phones or emails (Hann et al., 2008; Kumar, Zhang, and Luo, 2014; Goh, Hui, and Png, 2015), offline retail sales (Schneider et al., 2018), subscription services and various customer relationship management (CRM) programs (Conitzer, Taylor, and Wagman, 2012), online personalization services in computers and mobile devices (Chellappa and Shivendu, 2010), online search and e-commerce transactions (Bart et al., 2005), online social networks (Adjerid, Acquisti, and Loewenstein, 2018). Prior studies have also examined privacy topics related to finance and healthcare, such as crowd-funding (Burtch, Ghose, and Wattal, 2015), credit transactions, insurance (Garfinkel, Gopal, and Goes, 2002; Soleymanian, Weinberg, and Zhu, 2019), and healthcare (Garfinkel, Gopal, and Goes, 2002; Miller and Tucker, 2009; Miller and Tucker, 2017). As advertisers commonly target consumers by leveraging consumers' private information, the latest research has also investigated online, social media, and mobile advertising (Goldfarb and Tucker, 2011a; Conitzer, Taylor, and Wagman, 2012; Tucker, 2013; Gardete and Bart, 2018; Goldfarb and Tucker, 2011c; Rafieian and Yoganarasimhan, 2018; Goldfarb and Tucker, 2011b). Broadly speaking, any circumstances that involve customer databases would entail privacy concerns and needs for privacy protection (Garfinkel, Gopal, and Goes, 2002; Martin, Borah, and Palmatier, 2017; Muralidhar and Sarathy, 2006; Qian and Xie, 2015). As a result, even business-to-business (B2B) platforms incur privacy concerns and require effective strategies to address these concerns (Kalvenes and Basu, 2006). Nonetheless, as massive volumes of novel mobile location data emerge, which offer unparalleled opportunities to examine large populations' granular lifestyles and generate debatably more severe privacy concerns, more research is needed to quantify consumer privacy risks and devise privacy-preserving strategies.

Marketing research on consumer privacy falls into four main streams: consumer-, firm-, regulation-, and methodology- focused. Since our work is method focused, we will concisely review that here. The other three streams are discussed in the Appendix A.7.

Prior research has developed methodologies for regulatory bodies and firms to address privacy concerns. These methods fall under two broad categories: without data obfuscation and with as in our research. Without data obfuscation, these methods largely involve firms altering consumers' privacy perceptions, hence alleviating privacy concerns. Examples include altering the order of survey questions (Acquisti, John, and Loewenstein, 2012), revealing other consumers' attitudes towards privacy (Acquisti, John, and Loewenstein, 2012), altering the labels of privacy-protecting options (Adjerid, Acquisti, and Loewenstein, 2018), offering opt-in/out options (Kumar, Zhang, and Luo, 2014), granting enhanced privacy controls over, for instance, personally identifiable information (Tucker, 2013), allowing customers to remain anonymous with a cost (Conitzer, Taylor, and Wagman, 2012), or providing only aggregate instead of granular information (Sandıkçı et al., 2013). Consumers themselves may also take actions to preserve privacy, such as declining to answer certain survey questions, concealing addresses, or deflecting marketing solicitations (Hann et al., 2008). Globally, governments are

also providing regulatory protections, such as national do-no-call registries (Goh, Hui, and Png, 2015) and state genetic privacy laws (Miller and Tucker, 2017). Other methodologies, on the other hand, leverage obfuscation of original data or query outputs. The premise is that an entity, data collector in our setting, would perform data obfuscation to preserve consumer privacy before releasing the data to a third party, an advertiser for instance, while ensuring that the data remain usable. We will provide a more thorough survey of two sub-streams of this research based on the assumptions made when developing the relevant techniques (Clifton and Tassa, 2013).

1.2.2 Privacy-preserving Methodology I: Syntactic Models

The assumption of syntactic models is that the entity performing the obfuscation knows the type of threat that a stalker or malicious entity intends to perform on the shared data, and accordingly transforms the data to curtail that specific threat. The seminal work in this area was the concept of k -anonymity (Samarati and Sweeney, 1998) aimed at columnar data to ensure that given a column, there would be at least k records that take the same columnar value. This would ensure that a consumer is protected from a re-identification threat, that is, his/her record cannot be completely identified even if a stalker has some background knowledge, usually a subset of the consumer's columnar values.

Studies have shown that k -anonymity is NP hard and suffers from the curse of dimensionality (Meyerson and Williams, 2004). Variations of the concept of k -anonymity and heuristics to approximate k -anonymity have then been proposed (Aggarwal et al., 2005). Since k -anonymity primarily focuses on the re-identification threat, the method is susceptible to sensitive attribute inference when a stalker aims to only infer a particular column of a consumer rather than completely re-identify all the columnar values. ℓ -diversity (Machanavajjhala et al., 2006) and confidence bounding (Wang, Fung, and Philip, 2007) are proposed to address these shortcomings. ℓ -diversity accomplishes this by obfuscating data such that sensitive attributes are well represented for each consumer, while confidence bounding limits a stalker's confidence of inferring a sensitive value to a certain threshold. In the context of mobile location data, the above methodologies are shown to suffer from the curse of high dimensionality (Aggarwal, 2005), reducing an advertiser's utility. To address this, variations of k -anonymity, such as k^m -anonymity (Terrovitis, Mamoulis, and Kalnis, 2008) and complete k -anonymity (Bayardo and Agrawal, 2005), have been developed for high dimensional transaction data. However, these techniques only address re-identification threats and are still vulnerable to sensitive attribute inference. Further, while these techniques work well for high dimensional data, they do not explore obfuscation of temporal information crucial in extracting behavioral information from location data. Next, we will review some of the recent syntactic models proposed to obfuscate location data.

Extensions of the above traditional heuristics have been proposed to preserve privacy in simulated/synthetic location data (Chen et al., 2013; Terrovitis, Mamoulis, and Kalnis, 2008; Abul, Bonchi, and Nanni, 2008; Yarovoy et al., 2009), truck/car movements (Abul, Bonchi, and Nanni, 2008; Yarovoy

et al., 2009), or social media check-in data (Terrovitis et al., 2017; Yang, Qu, and Cudre-Mauroux, 2018). The seminal work by Abul, Bonchi, and Nanni, 2008 proposes (k, δ) anonymity to perform space generalization on location data. In other words, the trajectories are transformed so that k of them lie in a cylinder of the radius δ . A variation of k -anonymity is further developed for moving object databases (MOD) based on the assumption that MODs do not have a fixed set of quasi-identifiers (QIDs) (Yarovoy et al., 2009). The authors define the timestamps of the locations as QIDs and propose two obfuscation techniques based on space generalization. Few recent studies have explored variants of k and (k, δ) anonymity where the location trajectories are either distorted (Gao et al., 2014) or distorted and cloaked to a certain granularity (Huo et al., 2012; Chow and Mokbel, 2011; Hwang, Hsueh, and Chung, 2013). All these studies aim at protecting consumers from re-identification threats.

More recently, suppression techniques have garnered attention in obfuscating location data (Chen et al., 2013; Terrovitis, Mamoulis, and Kalnis, 2008; Terrovitis et al., 2017). For example, the seminal work by Terrovitis, Mamoulis, and Kalnis, 2008 presents a local suppression obfuscation technique assuming that a stalker has access to partial consumer trajectories, similar to the setting of the re-identification threat in our study. Built on this work, Terrovitis et al., 2017 further propose global suppression, separately from local suppression. Providing privacy guarantees against both identity and attribute linkage threats, Chen et al., 2013 develop $(K, C)_L$ privacy framework. The model requires three parameters from a data collector: a stalker's success probability thresholds in both types of threats and a parameter corresponding to a stalker's background knowledge. Instead of measuring the data utility with a rudimentary metric, the number of unique location points or frequent sequences preserved in the obfuscated data, as in Chen et al., 2013 and Terrovitis, Mamoulis, and Kalnis, 2008; Terrovitis et al., 2017, our research captures the data utility by tying it to a popular business objective of an advertiser – POI recommendation. In specific, we capture a consumer's historical preferences to locations and co-visitations over time from their location data, measure the utility of the data as the performance of a collaborative filtering and a time-aware POI recommendation technique (Yuan et al., 2013; Yuan, Cong, and Sun, 2014).

1.2.3 Privacy-preserving Methodology II: Differentially Private Algorithms

This sub-stream of research is based on the concept of ϵ -differential privacy (Dwork and Lei, 2009). Differentially private algorithms guarantee that a stalker would make the same inference from the shared data whether or not a focal individual is included in the data. Unlike syntactic models, they are not limited to a specific type of threats, thus presenting a much stronger privacy notion. The obfuscation performed on the data usually involves perturbation, that is, adding a noise to the data before sharing them (Muralidhar and Sarathy, 2006). Another related method is data shuffling, which is usually performed across rows or columns, such as replacing a subset of a consumer's record with another consumer's record to minimize privacy risks. Various studies have leveraged perturbation, data shuffling, or a combination of them (Qian and Xie, 2015). For instance, Garfinkel, Gopal, and Goes,

2002 perturb the answer of a database query to generate the correct answer probabilistically or deterministically embedded in the range of the perturbed answers. Muralidhar and Sarathy, 2006 employ data shuffling for confidential numerical data where the values of the confidential variables are shuffled among observations, while preserving a high level of data utility and minimizing the risk of disclosure. Schneider et al., 2018 develop a Bayesian probability model to produce synthetic data. Besides perturbation and data shuffling, public key encryption, digital certificate, and blinded signatures are also common privacy-friendly tools (Kalvenes and Basu, 2006). All of the above methods focus on columnar data.

In the context of location data, while data querying has been studied (Riboni and Bettini, 2012; Pelekis et al., 2011; Hwang, Hsueh, and Chung, 2013; Guo et al., 2015; Wernke et al., 2014), the literature on data sharing is sparse. A few techniques have been developed to generate synthetic trajectories from a series of differentially private queries (He et al., 2015; Chen, Acs, and Castelluccia, 2012). The utility of the data preserved while generating these trajectories usually involves summary statistics, such as the number of unique locations or frequent location patterns. Moreover, owing to the stronger theoretical guarantees to be met, these techniques have been empirically shown to not preserve the truthfulness of the location data, hence hindering advertisers' abilities to perform sophisticated data mining tasks (Terrovitis et al., 2017). In our research, the consumers have explicitly opted in to share their location data with the data collector and advertisers in exchange for personalized offers. So we take the route of syntactic models that are more likely to result in a higher data utility to an advertiser. We assume that a data collector has reasonable knowledge about the type of privacy threats that a consumer could be exposed to. To minimize the privacy threats, we propose an obfuscation scheme based on suppression that also ensures sufficient utility of the obfuscated location data to an advertiser.

Our study distinguishes itself from the prior research along several dimensions. We detail these differences with closely related works in Table 1.1. Methodologically, we quantify the privacy risk at a consumer level, instead of an aggregate or location level as in the prior literature (Terrovitis, Mamoulis, and Kalnis, 2008; Terrovitis et al., 2017). We also measure the utility of the location data in the context of real-world business applications, such as POI recommendation, instead of using the aggregate or rudimentary metrics from the literature, such as the number of unique locations or frequent sequences (He et al., 2015). From the standpoint of practical applicability, the proposed framework requires merely one parsimonious input, the number of locations already known to a stalker (Section 1.4.1). Therefore, it is intuitive and interpretable to the data collector or any manager. We also provide a data collector with multiple options of the risk-utility trade-off. Finally, most prior studies have validated their recommendations only on synthetic data (Chen et al., 2013; Terrovitis, Mamoulis, and Kalnis, 2008; Abul, Bonchi, and Nanni, 2008; Yarovoy et al., 2009), vehicle movements (Abul, Bonchi, and Nanni, 2008; Yarovoy et al., 2009), or social media check-ins (Terrovitis et al., 2017; Yang, Qu, and Cudre-Mauroux, 2018) with various data limitations described earlier, such as accuracy or representativeness. In contrast, we validate our proposed framework on granular mobile location data

Paper	Consumer Privacy Threats	Advertiser Utility	Obfuscation scheme	Empirical Data	Consumer level quantification?	Consumer level obfuscation?
Abul, Bonchi, and Nanni, 2008	(k, δ) -anonymity	Deviation from true trajectories	Distortion / Cloaking	Simulated data	\times	\times
Yarovoy et al., 2009	(k) -anonymity	Information loss	Distortion	Car trajectories	\times	\times
Terrovitis, Mamoulis, and Kalnis, 2008	Re-identification threat	Information loss	Suppression	Transaction Data	\times	\times
Terrovitis et al., 2017	Re-identification threat	Frequent sequences	Split and suppression	Social network trajectories	\times	\times
Chen et al., 2013	Re-identification threat Sensitive Attribute threat	Frequent sequences	Suppression	Simulated Data	\times	\times
Gao et al., 2014	k -anonymity	Information loss	Distortion	Simulated data	\times	\checkmark
Xue et al., 2013	Home address leakage	Destination prediction	Synthesis	Taxi trajectories	\times	\checkmark
Huo et al., 2012	(k, δ) anonymity	Information loss	Split and distort	155 consumer trajectories	\times	\times
Pelekis et al., 2011	Sensitive location protection	Streaming k-NN queries	Distortion and Synthesis	Simulated data	\times	\checkmark
Chow and Mokbel, 2011	k -anonymity	Deviation from true trajectories	Distortion / Cloaking	NA	\times	\times
Hwang, Hsueh, and Chung, 2013	r -anonymity	Number of consumers in a region	Distortion / Cloaking	Taxi trajectories	\times	\checkmark
Riboni and Bettini, 2012	ϵ -DP	Streaming POIs	Distortion / Cloaking	NA	\times	\checkmark
He et al., 2015	ϵ -DP	Frequent patterns	Synthesis	Taxi and network trajectories	\times	\times
Chen, Acs, and Castelluccia, 2012	ϵ -DP	Frequent patterns	Synthesis	Metro trajectories	\times	\times
Proposed work	Re-identification threat Sensitive Attribute threat Flexibility to incorporate other estimates of consumer risk	POI prediction Activity Prediction Flexibility to incorporate other estimates of consumer utility.	Suppression	40k consumer trajectories	\checkmark	\checkmark

TABLE 1.1: Comparison of Proposed Method to Relevant Literature in Privacy Preserving Location Data Sharing

from a large population over time.

1.2.4 Location-based Mobile Marketing

Finally, our work is related to the research on location-based mobile marketing. Using randomized field experiments, researchers have demonstrated that mobile advertisements based on the location and time information can significantly increase consumers' likelihood of redeeming geo-targeted mobile coupons (Fang et al., 2015; Molitor et al., 2019; Fong, Fang, and Luo, 2015b; Luo et al., 2014). In our framework, we measure the utility of the location data to an advertiser by considering a popular business application, POI recommendation. Identifying the next location most likely visited by a consumer based on his or her prior trajectories is crucial to perform behavioral targeting. Ghose, Li, and Liu, 2018 design a POI-based mobile recommendation based on similarities of consumers' mobile trajectories and demonstrate that such a strategy can lead to a significant improvement in a retailer's incremental revenues. Other recent studies have revealed that understanding consumers' hyper-context, for example, the crowdedness of their immediate environment (Andrews et al., 2016), weather (Li et al., 2017), or the competitive choices (Fong, Fang, and Luo, 2015a; Dubé et al., 2017), is also critical to marketers' evaluations of the effectiveness of mobile marketing. Another group of studies have further examined consumers' perceptions and attitudes toward location-based mobile marketing (Bruner and Kumar, 2007; Xu, 2006). In the next section, we will describe the mobile location data under analysis.

1.3 Data

We partner with a leading U.S. data collector that aggregates location data across hundreds of commonly used mobile apps, from news, weather, map, to fitness. The data cover one-quarter of the U.S. population across Android and iOS operating systems.

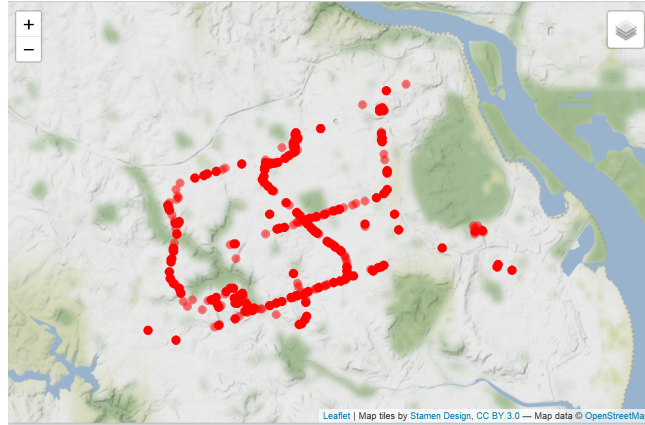


FIGURE 1.2: An example of a consumer's footprints with 732 unique locations over the five-week sample period

Description	Mean (S.D.)	Min (Max)
Number of locations per person	23.47 (50.26)	2 (1104)
Number of unique locations per person	14.25 (38.12)	2 (963)
Overall duration (in hours)	272.97 (278.25)	0.05 (759.27)
Duration at each location (minutes)	27.96 (45.99)	1.6 (359.23)
Distance between locations (in km)	1.89 (3.89)	0.02 (75.49)

TABLE 1.2: Summary statistics of the location data sample under analysis

The data sample under analysis covers a major mid-Atlantic metropolitan region in the U.S. Figure 1.2 displays the region's map (blurred on purpose due to a confidentiality agreement) and an example of a consumer's footprints with 732 unique locations visited during our five-week sampling period between September and October, 2018. The entire sample includes 940,000 locations from 40,012 consumers. Each row of the data corresponds to a location recorded for a consumer and contains information about

- Consumer ID: a unique identifier of each consumer;
- Platform ID: an identifier of the consumer's mobile operating system (Android or iOS);
- Latitude and longitude (i.e., geo-coordinates) of the location visited;
- Timestamp: the beginning time at the location.
- Time spent: The amount of time spent at the location.

We randomly sample 50% of all consumers in the data (20,000 consumers) and all their location data for training and cross-validating our machine learning models (Section 1.5 and Appendix A.6). Based on the models and parameters trained, we then conduct the focal analysis using the remaining 50% of the data. Table 1.2 displays the summary statistics of the data. On average, a consumer visited from 2 to 963 unique locations tracked by the data. To reduce smartphone battery drainage, data redundancy, and storage cost, each consumer's smartphone is pinged frequently, but only recorded a location when there is a substantial change in the geo-coordinates. The average duration at each location is 27.96 minutes. The average overall duration, measured as the difference of a consumer's last and first time stamp is 272 hours (≈ 1.6 weeks). And the Euclidean distance between any two consecutively tracked locations is 1.89 km on average after converting the locations' latitudes and longitudes to the Universal Transverse Mercator (UTM) coordinates.

The literature on privacy-preserving sharing of location data has tested the methodologies on simulated data (Chen et al., 2013; Terrovitis, Mamoulis, and Kalnis, 2008; Abul, Bonchi, and Nanni, 2008; Yarovoy et al., 2009), vehicle movements (Abul, Bonchi, and Nanni, 2008; Yarovoy et al., 2009), or social media check-ins (Terrovitis et al., 2017; Yang, Qu, and Cudre-Mauroux, 2018), also only over a short period, such as 24 hours. We make an initial effort to develop a privacy-preserving framework for, and validate it on, a real-world human physical movement data across a large population. Such data are automatically tracked in real time by mobile devices, often via wifi, beacons, and GPS etc. multi-technology multilateration with an accuracy radius of merely 20 meters. They are thus much more precise than cell tower tracking that often has an accuracy radius of a few kilometers, social media geo-tags known for its sparsity and inaccuracy, or consumers' self check-ins that rely on consumers' manual labor and willingness to check-in at any location. The mobile location data under our study are also more representative of the general population than taxi or public transportation data, hence much more valuable to advertisers and other data users. On the other hand, these data's massive scale and high dimensionality, in our case nearly one million mobile location over just five weeks from one metropolitan region, also entail unique challenges as discussed earlier, hence imminent needs to develop new privacy-preserving frameworks that can address these challenges.

1.4 Methodology

The proposed framework enables a location data collector to share data in a privacy preserving manner while ensuring sufficient utility to an advertiser from the shared data. Consistent with the premise of syntactic models, a data collector has some knowledge about the types of potential privacy threats (Clifton and Tassa, 2013). While the proposed framework accommodates various types of privacy threats, we illustrate two commonly encountered types - sensitive attribute inference and re-identification threat. We will introduce the notations first and then formulate the privacy preservation in the context of the location data.

Definition 1 A trajectory T_i of a consumer i is defined as a temporally ordered set of tuples $T_i = \{(l_i^1, t_i^1), \dots, (l_i^{n_i}, t_i^{n_i})\}$, where $l_i^k = (x_i^k, y_i^k)$ is a location k visited by consumer i with geo-coordinates (i.e., a pair of longitude and latitude) x_i^k and y_i^k , t_i^k is the corresponding timestamp, and n_i is the total number of locations tracked of consumer i .

Problem Formulation. We frame the problem of preserving privacy in location data at a consumer level. Let r_i denote a consumer i 's privacy risk associated with trajectory T_i for a specific type of privacy threat, and u_i the advertiser's utility from leveraging consumer i 's trajectory. A data collector aims to find a transformation $T_i \rightarrow \mathcal{P}(T_i)$, where $\mathcal{P}(T_i)$ is consumer i 's obfuscated trajectory that the data collector shares with an advertiser by minimizing r_i while maintaining u_i . The transformation is based on suppressing the locations in T_i given two suppression parameters. One is \vec{s}_i , the suppression weight corresponding to each unique location in T_i . It is assigned based on various measures of the informativeness of each location, such as the consumer's frequency, recency, and time spent at each location. The more informative a location is, the more likely it is suppressed. The other is z_i , the suppression score for consumer i , which controls the number of locations in T_i to be suppressed. It is assigned based on the consumer's privacy risk. The higher the risk for consumer i , the more locations are suppressed in T_i . Both parameters contribute to the final suppression probabilities assigned to each location in T_i . In Section 1.4.3, we will detail a structured grid search to fine-tune these two parameters, which do not need to be input by a data collector. The corresponding risk and utility of the obfuscated trajectory $\mathcal{P}(T_i; \{\vec{s}_i, z_i\})$ are functions of the two suppression parameters,

$$\begin{aligned} r_i &= \mathcal{PR}(T_i; \{\vec{s}_i, z_i\}) \\ u_i &= \mathbf{U}(T_i; \{\vec{s}_i, z_i\}), \end{aligned}$$

where $\mathcal{PR}(\cdot)$ and $\mathbf{U}(\cdot)$ depend on the type of privacy threat and business objective of the advertiser, respectively.

Overall, for a set of N consumers' trajectories $T = \{T_1, \dots, T_N\}$, the data collector aims to find a transformation of T , $T \rightarrow \mathcal{P}(T; \{\vec{s}_i, z_i\}_{i=1}^N)$, to produce the obfuscated trajectories to be shared with the advertiser that minimize the expected privacy risk $E(r_i)$ across all consumers while maintaining the expected data utility $E(u_i)$ to the advertiser. Consistent with our focal research questions and overview of the three components of the proposed framework (Fig. 1.1), we further break down the data collector's problem into three sub-problems below. The first two pertain to the estimation of u_i and r_i based on \mathcal{PR} and \mathbf{U} , respectively; and the third is to identify the suppression parameters $\{\vec{s}_i, z_i\}$.

Problem 1 Quantification of Consumer's Privacy Risk: Given the consumers' trajectories T and a privacy threat \mathcal{PR} , we quantify each consumer's risk $\{r_1, \dots, r_N\}$, where each $r_i \in [0, 1]$ indicates the stalker's success rate in inferring the private information from consumer i 's trajectory T_i .

Problem 2 Quantification of Advertiser's Utility: Given the consumers' trajectories T and a business objective \mathbf{U} , we quantify each trajectory's utility to an advertiser $\{u_1, \dots, u_N\}$.

Problem 3 Obfuscation Scheme for Data Collector: Given consumer trajectories T and their corresponding risks, for an advertiser's business objective \mathbf{U} , we identify an obfuscation scheme $T \rightarrow \mathcal{P}(T; \{\vec{s}_i, z_i\}_{i=1}^N)$ to balance the average risk and utility across consumers.

Next, we will illustrate the quantification of two classes of privacy risks in Section 1.4.1 and quantification of the data's utility to an advertiser in one business application of POI recommendation in Section 1.4.2. Finally, in Section 1.4.3, we will propose an obfuscation scheme that provides a balance between the privacy risks and data utility.

1.4.1 Quantification of Consumer's Privacy Risk

The first step of the proposed framework is quantifying each consumer's privacy risk. To accomplish this, we simulate a stalker's actions and assign its success rate in obtaining a consumer's sensitive information as the consumer's privacy risk. Privacy threats could range from using simple heuristics, such as querying the consumers' trajectories, to leveraging more robust machine learning heuristics to predict consumers' sensitive attributes (Li, Shirani-Mehr, and Yang, 2007; Yang, Qu, and Cudre-Mauroux, 2018). In our framework, we consider both simple and sophisticated heuristics. Specifically, we will examine two types of the most commonly encountered stalker threats. The first type is "sensitive attribute inference", where a stalker could employ robust machine learning heuristics to infer sensitive information, such as home address (Yang, Qu, and Cudre-Mauroux, 2018). The second type is "re-identification threat", where a stalker aims to infer a consumer's complete set of locations T_i , that is, identify consumer i , from the published trajectories $\mathcal{P}(T)$ (Pellungrini et al., 2018). With some background knowledge, such as a subset of a consumer's locations $\bar{T}_i \in T_i$, a stalker could query the published trajectories $\mathcal{P}(T)$ to identify a subset of J consumers who have visited all locations in \bar{T}_i . A lower J indicates a higher re-identification risk.

To replicate a stalker's adversarial actions and assess each consumer's privacy risks, we extract a comprehensive set of features from the trajectories – $\mathcal{F}(T)$ ⁴ to capture consumer mobility patterns and consumer-location, consumer-consumer affinities (Gonzalez, Hidalgo, and Barabasi, 2008; Eagle and Pentland, 2009; Williams et al., 2015; Pappalardo, Rinzivillo, and Simini, 2016; Ashbrook and Starner, 2003; Zheng, Xie, and Ma, 2010; Wang et al., 2011). These extracted features, as we will see later in Section 1.5.1, will also help a data collector interpret which features contribute the most to the privacy risks, gain insights on possible obfuscation schemes, and quantify and interpret the data utility to an advertiser. A detailed description of these features is presented next.

Trajectory Feature Extraction.

To replicate a stalker's adversarial actions and assess each consumer's privacy risks, we extract a comprehensive set of features from the trajectories

⁴To simplify the notation, we use $\mathcal{F}(T)$ to refer to $\mathcal{F}(\mathcal{P}(T))$. As we will see later in Section 1.4.3, both $\mathcal{P}(T)$ and T are a set of trajectories as defined Def. 1. Hence, any operation (\mathcal{F} here) performed on T is applicable to $\mathcal{P}(T)$ as well.

examined by the literature, $\mathcal{F}(T)$ ⁵ (Gonzalez, Hidalgo, and Barabasi, 2008; Eagle and Pentland, 2009; Williams et al., 2015; Pappalardo, Rinzivillo, and Simini, 2016; Ashbrook and Starner, 2003; Zheng, Xie, and Ma, 2010; Wang et al., 2011).

1. **Consumer Mobility:** This set of features captures a consumer’s aggregate mobility patterns based on the locations visited in T_i , such as the consumer’s frequency to, time spent at (Pappalardo, Rinzivillo, and Simini, 2016), and distance traveled to a location (Williams et al., 2015). We also compute other richer mobility features, such as entropy (Eagle and Pentland, 2009) and radius of gyration (Gonzalez, Hidalgo, and Barabasi, 2008). A detailed description of these features is listed in Table 1.3.
2. **Consumer-Location Affinity:** Leveraging the literature on learning significant locations from predicting movements across trajectories (Ashbrook and Starner, 2003; Zheng, Xie, and Ma, 2010), we build three arguably most straightforward consumer-location tensors: the frequency to, time spent at, and total distance traveled from the immediate prior location to each location by a consumer at a weekly level. Each of these three tensors is of order three—consumer by unique location by week. We then extract consumer specific, lower dimensional representations by performing a higher order singular value decomposition (HOSVD) on the three tensors separately (De Lathauwer, De Moor, and Vandewalle, 2000). HOSVD is typically applied to extract features from multivariate data with temporal and spatial dimensions similar to ours (Fanaee-T and Gama, 2015). Since the tensors are populated over the locations visited by these consumers, the extracted features would effectively capture the affinity of the consumers to significant locations.
3. **Consumer-Consumer Affinity:** Prior studies have also predicted consumer network or social links based on trajectories (Wang et al., 2011). We thus quantify the consumers’ co-location behaviors by building consumer-consumer affinity tensors based on the locations that the consumers share at a weekly level. Each tensor would of order three —consumer by consumer by week. We populate three such tensors with the average frequency to, total time spent at, and distance traveled to each co-visited location within a week, respectively. Next, we perform a HOSVD on each of these three tensors to extract the consumer specific low dimensional representations indicative of the affinity to other consumers. The incremental benefit of the affinity features is discussed in Appendix A.5.

Stylized Example. We illustrate the above consumer-location and consumer-consumer affinity features using a stylized example. Consider three consumer trajectories as defined in Definition 1: $T_1 = \{(A, 1), (B, 1), (A, 2), (A, 2)\}$, $T_2 = \{(C, 1), (A, 1), (A, 1)\}$, $T_3 = \{(D, 1), (B, 1), (C, 2)\}$, where A, B, C, D are location identifiers and the granularity of the timestamps is at a weekly level. That is, $T = \{T_1, T_2, T_3\}$ reveals that these three consumers visited four

⁵To simplify the notation, we use $\mathcal{F}(T)$ to refer to $\mathcal{F}(\mathcal{P}(T))$. As we will see later in Section 1.4.3, both $\mathcal{P}(T)$ and T are a set of trajectories as defined Def. 1. Hence, any operation (\mathcal{F} here) performed on T is applicable to $\mathcal{P}(T)$ as well.

Feature	Description
average_locations	Number of locations in T_i averaged weekly.
average_ulocations	Number of unique locations in T_i averaged weekly.
average_distance	Distance travelled by a consumer to visit locations in T_i , averaged weekly.
average_dwell	Time spent at locations in T_i averaged weekly.
avg_max_distance (Williams et al., 2015)	Average of the maximum distance travelled by a consumer each week.
freq_rog, time_rog, dist_rog (Gonzalez, Hidalgo, and Barabasi, 2008)	Radius of gyration is the characteristic distance traveled by an individual. $rog_i = \sqrt{\frac{1}{ T_i } \sum_{j=1}^{ T_i } w_{ij} (l_{ij} - l_{cm}^i)^2}$ $l_{cm}^i = \frac{1}{ T_i } \sum_{j=1}^{ T_i } l_{ij}$ l_{ij} are the geographical coordinates l_{cm}^i is the center of mass of the consumer w_{ij} are weights obtained based on frequency, time & distance w.r.t to l_{ij}
freq_entropy, time_entropy, dist_entropy (Eagle and Pentland, 2009)	Mobility entropy measures the predictability of consumer trajectory. $E_i = -\sum_{j=1}^{ T_i } p_{ij} \log_2 p_{ij}$, p_{ij} computed from w_{ij} for time, frequency & distance.

TABLE 1.3: Description of consumer mobility features

unique locations over a period of two weeks. Each of the three consumer-location tensors discussed above would be of size $[3 \times 4 \times 2]$ for the 3 consumers, 4 unique locations, and 2 weeks. For instance, the frequency matrix of the first consumer with T_1 is $\begin{pmatrix} 1 & 1 & 0 & 0 \\ 2 & 0 & 0 & 0 \end{pmatrix}$, where the rows and columns correspond to the 2 weeks and 4 unique locations, respectively, and each entry in the matrix captures the number of times that this consumer visited each of the four locations during that week. Each of the three consumer-consumer location tensors described above would be of size $[3 \times 3 \times 2]$ for the 3 consumers by 3 consumers by 2 weeks. For instance, the frequency matrix for the first consumer with T_1 would be $\begin{pmatrix} 1 & \frac{(1+2)}{2} & \frac{(1+1)}{2} \\ 1 & 0 & 0 \end{pmatrix}$, where the rows and columns correspond to weeks and the consumer pairs 1-1, 1-2, and 1-3. Each entry in this matrix is the average frequency of the co-visited locations within each consumer pair. For instance, during week 1, $(A, 1)$ was co-visited by consumers 1 and 2, and $(B, 1)$ by consumers 1 and 3. The time and distance tensors are similarly constructed. We then perform a HOSVD on these tensors separately and use the first five principal components that capture a majority of the variance. Hence, for each consumer and tensor, we have five lower dimensional representations that capture the corresponding consumer-location and consumer-consumer affinities.

Next, we imitate how a stalker would use the extracted features from the published trajectories to orchestrate privacy threats.

Sensitive Attribute Inference.

Leveraging the published trajectories $\mathcal{P}(T)$ and extracted features, a stalker could infer various sensitive attributes, thus posing a privacy threat (Li, Shirani-Mehr, and Yang, 2007). We train a supervised model \mathbf{M}_{proxy} with the extracted features as a proxy for the stalker’s model \mathbf{M} to infer the sensitive attributes (Yang, Qu, and Cudre-Mauroux, 2018). Each consumer’s risk is quantified as the certainty of identifying a sensitive attribute from the consumer’s published trajectory using \mathbf{M}_{proxy} . We illustrate the method by inferring two sensitive attributes, home address (discussed in Section 1.5) and mobile operating system (deferred to Appendix A.8).

Specifically, we enlist Random Forest as \mathbf{M}_{proxy} in light of its flexibility in handling regression and classification tasks, and its competitive performance across a wide range of supervised learning algorithms (Breiman, 2001; Liaw, Wiener, et al., 2002). For each sensitive attribute, we learn a Random Forest using the extracted features⁶. The risk is then calculated as the certainty of \mathbf{M}_{proxy} in identifying the corresponding sensitive attribute, that is, the probability of correctly identifying the attribute in classification, or negative root mean square error in regression. We also perform a 0-1 normalization in regression such that $r_i \in [0, 1]$.

Re-identification Threat.

Adapting the risk notion that a stalker is able to identify a consumer and associate the consumer with a record in the published data (Samarati, 2001; Samarati and Sweeney, 1998), we define re-identification threat in the context of location data. Here, a stalker tries to re-identify all locations visited by a consumer based on some prior knowledge of an (often small) subset of locations visited by the consumer, such as employer address from a membership registration form. Formally, this problem can be defined as follows:

Definition 2 *Given the published trajectories $\mathcal{P}(T)$ and a subset of consumer i ’s trajectory $\bar{T}_i \subseteq T_i$, the stalker aims to identify T_i from $\mathcal{P}(T)$.*

Since a data collector does not know consumer i under threat or the subset locations \bar{T}_i a-priori, to quantify the consumer’s risk r_i , the data collector would need to account for all $\binom{|T_i|}{|\bar{T}_i|}$ possible subsets of T_i , where $|T_i|$ is the total number of unique locations visited by a consumer i . For each such subset, the probability of a consumer being identified is $\frac{1}{J}$, where J denotes the number of all consumers among N who have visited all locations in \bar{T}_i . If no such consumer exists other than i , then the probability of identifying consumer i would be 1 for the subset considered. We quantify a consumer’s re-identification risk as the maximum of these probabilities over all such subsets. To reduce the computational complexity of estimating re-identification risk, we employ a speed-up heuristic leveraging a recent study (Pellungrini et al., 2018). This is discussed in Appendix A.4

⁶We have also compared Random Forest with a number of tree-based and boosting classification methods – xGBoost (Chen and Guestrin, 2016), Conditional inference trees (Hothorn, Hornik, and Zeileis, 2015), Adaboost (Hastie et al., 2009); and found that Random Forest provides the best out-of-sample performance.

A Stylized Example. Three consumers' trajectories over a two-week period, $T_1 = \{(A,1), (B,1), (C,2), (C,2)\}$, $T_2 = \{(A,1), (B,1), (A,2)\}$, $T_3 = \{(A,1), (B,1), (C,2)\}$, suggest that all three consumers visited the location subset (A, B), two of them (consumers 1 and 3) visited (B, C), and two (consumers 1 and 3) visited (A, C). Then given each of these location subsets, the corresponding probabilities of identifying consumer 1 are $\{\frac{1}{3}, \frac{1}{2}, \frac{1}{2}\}$, resulting in consumer 1's re-identification risk as $\max(\frac{1}{3}, \frac{1}{2}, \frac{1}{2}) = \frac{1}{2}$. The overall intuition behind the re-identification risk is that given a similar number of unique locations visited across consumers, a person who visits more unique locations not visited by others would have a higher re-identification risk.

1.4.2 Quantification of Advertiser's Utility

Having quantified each consumer's privacy risks associated with the two commonly encountered privacy threats (our research question 1), we next examine the utility that an advertiser would derive from the published trajectories (research question 2). The behavior-rich nature of the location data enables advertisers to derive great insights and perform various targeted marketing activities to reap monetary benefits. In this work, we consider a popular business application, POI recommendation (Ashbrook and Starner, 2003). The underlying idea is to leverage the historical consumer preferences revealed in the trajectories to predict the locations that a consumer is most likely to visit in the future. This would enable an advertiser to target the consumer with relevant, contextualized marketing messages (Ghose, Li, and Liu, 2018). To this end, we quantify an advertiser's utility by learning a recommendation model. Intuitively, more accurate POI predictions will render better targeting and thus a higher utility for the advertiser. Hence, we quantify u_i , the utility of consumer i 's trajectory, as the predictive accuracy of the recommendation model.

Most recommendation models leverage collaborative filtering to identify other consumers with similar historical preferences to infer the focal consumer's preference (Bobadilla et al., 2011). This idea is consistent with human social behavior: people tend to account for their acquaintances' tastes, opinions, and experiences when making own decisions. We thus imitate an advertiser's use of the location data for POI recommendation and compare a number of recommendation models examined in the literature (Appendix A.6). We focus the following discussion on the best performing nearest neighborhood (NN) based learning technique. Simply put, the main idea of NN is to identify the m consumers most similar to the focal consumer, namely m neighbors, and utilize their locations to predict the focal consumer's future locations. The similarity is computed based on the visited locations that reveal each consumer's preference by leveraging the set of features extracted from the published trajectories described in Section 1.4.1. To find the m most similar consumers, we compute the cosine similarity between two consumers' features $\mathcal{F}(T_i)$ and $\mathcal{F}(T_j)$:

$$\text{sim}(\mathcal{F}(T_i), \mathcal{F}(T_j)) = \frac{\mathcal{F}(T_i) \cdot \mathcal{F}(T_j)}{\|\mathcal{F}(T_i)\| \|\mathcal{F}(T_j)\|} \quad (1.1)$$

After identifying the m most similar consumers to a consumer i , denoted as M_i , we aggregate and rank the unique locations visited by M_i based on a combination of visit frequency and these m consumers' similarities to consumer i . Specifically, for each consumer $j \in M_i$, location $l \in T_j$, let f_j^l denote the number of times that consumer j visited location l , then the rank of a location l for consumer j is determined by:

$$o_{ij}^l = \sum_{l=1}^{|T_j|} \frac{f_j^l}{\sum_l f_j^l} \text{sim}(\mathcal{F}(T_i), \mathcal{F}(T_j)) \quad (1.2)$$

In the above equation, $\frac{f_j^l}{\sum_l f_j^l}$ is the normalized visit frequency at a consumer level for a location. Intuitively, Equation 1.2 ensures that an individual i is most likely to visit the most frequently visited location of the most similar consumer. We further aggregate o_{ij}^l across all the consumers who visited the location l in M_i by computing the mean of o_{ij}^l :

$$o_i^l = \frac{1}{\sum_{j=1}^{|M_i|} \mathbf{1}(l \in T_j)} \sum_{j=1}^{|M_i|} \mathbf{1}(l \in T_j) \cdot o_{ij}^l \quad (1.3)$$

where $\mathbf{1}(j \in T_j) = 1$ if consumer j has visited location l and zero otherwise. The higher the value of o_i^l , the more likely that a consumer i visits location l in the future. The next k locations (ordered by time) most likely visited by consumer i hence correspond to the top k such ranked locations. The utility of consumer i 's trajectory T_i to the advertiser is then measured as the predictive accuracy of the recommendation model for the different values of k , measured by the widely used information retrieval metrics that assess the quality of the recommendations: Average Precision at k ($AP@k$ or AP_i^k) and Average Recall at k ($AR@k$ or AR_i^k) (Yang, Qu, and Cudre-Mauroux, 2018). Specifically, let $L_i = \{l_i^1, l_i^2, \dots, l_i^k\}$ be the actual next k' locations visited by consumer i and $\bar{L}_i = \{\bar{l}_i^1, \bar{l}_i^2, \dots, \bar{l}_i^k\}$ be the top k locations predicted by the NN recommendation model as described above. Then AP_i^k and AR_i^k are:

$$AP_i^k = \frac{1}{|L_i \cap \bar{L}_i|} \sum_{j=1}^k \frac{|L_{1:j} \cap \bar{L}_{1:j}|}{|L_{1:j}|} \quad (1.4)$$

$$AR_i^k = \frac{1}{|L_i \cap \bar{L}_i|} \sum_{j=1}^k \frac{|L_{1:j} \cap \bar{L}_{1:j}|}{|L_i|} \quad (1.5)$$

The intuition is that AP_i^k measures the proportion of the recommended locations that are relevant, while AR_i^k measures the proportion of relevant locations that are recommended. Then the expected utility of all consumers' trajectories to the advertiser $E(u_i)$ is calculated as $MAP@k$ and $MAR@k$, i.e., the mean AP_i^k and mean AR_i^k , respectively, across all consumers. Also,

the parameter m (number of the most similar neighbors) is selected by performing a five-fold cross-validation aimed at maximizing the accuracy of the recommendations (details in Section 1.5.2), a technique commonly used in the statistical learning literature to ensure a good out-of-sample performance (Friedman, Hastie, and Tibshirani, 2001).

1.4.3 Obfuscation Scheme

The last step in our framework is to address the third research question – devising an obfuscation scheme for the data collector that would balance the privacy risks to the consumers and the utility of the published trajectories to the advertiser. As discussed earlier, given the unique properties of trajectory data, such as high dimensionality, sparsity, and sequentiality, employing the traditional obfuscation techniques proposed for relational data, such as k -anonymity (Samarati and Sweeney, 1998), ℓ -diversity (Machanavajjhala et al., 2006), and confidence-bounding (Wang, Fung, and Philip, 2007) would be computationally prohibitive and significantly reduce the utility of the resulting obfuscated data (Aggarwal, 2005). On the other hand, those techniques devised specifically for trajectory data are often complex for a data collector to interpret and apply in practice. For instance, the $(K, C)_L$ privacy framework (Chen et al., 2013) requires multiple parameter inputs from a data collector, including the threshold of the stalker’s success probability and the stalker’s background knowledge in each type of threat. LSUP (Terrovitis et al., 2017) requires similar inputs. Given the complex nature of such heuristics, setting these parameters and interpreting the resulting obfuscations for practical purposes is non-trivial. Moreover, the current techniques do not provide the flexibility for a data collector to choose among multiple obfuscation schemes. Addressing these critical challenges, we develop $T \rightarrow P(T, \{\vec{s}_i, z_i\}_{i=1}^N)$, a personalized consumer-level suppression technique that is interpretable to the data collector. It requires no input parameter for the sensitive attribute inference and merely one input parameter for the re-identification threat – the number of a consumer’s locations already known to the stalker $|\bar{T}_i|$. Furthermore, the data collector will enjoy the flexibility of choosing among multiple interpretable obfuscations for each type of privacy threat.

In our obfuscation scheme, a consumer’s trajectory T_i is suppressed based on two consumer-specific suppression parameters $\{\vec{s}_i, z_i\}$. As described earlier, the suppression score z_i controls the number of locations to be suppressed in a consumer i ’s trajectory T_i , and the suppression weights \vec{s}_i denote the likelihood for each unique location to be suppressed. A naive approach to identify $\{\vec{s}_i, z_i\}$ that balance the risk and utility is to search over a random grid of positive values of \vec{s}_i and z_i . However, this would be computationally inefficient, contingent on the grid of values chosen, and potentially resulting in no parameters that could satisfactorily balance the risk and utility and hence requires a more sophisticated grid search.

A more structured approach to identify the parameters would be to consider a grid that ensures reduction in consumer’s risk and assesses the corresponding reduction in utility to pick a specification that satisfactorily balances the risk-utility trade-off. Intuitively, more locations suppressed would

mean lower risks to the consumers and lower utility to the advertiser; and in the extreme scenario of no trajectories published, both risk and utility would be zero. Also, to ensure similar risk reduction for a high-risk and a low-risk consumer, the number of locations suppressed would need to be proportional to the consumer's privacy risk r_i , that is, $z_i = r_i \times p$, where $p \in [0, 1]$ is a grid parameter.

While z_i ensures that the number of locations suppressed is proportional to the consumer's risk r_i , to further limit the information available to perform a stalker threat, the more informative locations within T_i would need to be suppressed with higher probabilities. Since the informativeness is related to the possible features that can be extracted by a stalker from T_i (Section 1.4.1), we assign the suppression weights \vec{s}_i based on the key features capturing the informativeness - frequency, recency, or time spent at each location. To exemplify, let $L_i = \{l_i^1, l_i^2, \dots, l_i^{k_i}\}$, be the unique locations in $T_i = \{(l_i^1, t_i^1), \dots, (l_i^{k_i}, t_i^{k_i})\}$, $k_i \leq n_i$. Then the weights based on the corresponding frequencies $\{f_i^1, f_i^2, \dots, f_i^{k_i}\}$ are $\vec{s}_i = \left\{ \frac{f_i^1}{\sum_{j=1}^{k_i} f_i^j}, \frac{f_i^2}{\sum_{j=1}^{k_i} f_i^j}, \dots, \frac{f_i^{k_i}}{\sum_{j=1}^{k_i} f_i^j} \right\}$.

Combining the two parameters $\{\vec{s}_i, z_i\}$ described above, we can calculate the suppression probability of each unique location in T_i . Then the unique locations are independently suppressed with Bernoulli trials given the following probabilities:

$$z_i + z_i \times s_i^1, z_i + z_i \times s_i^2, \dots, z_i + z_i \times s_i^{k_i} \quad (1.6)$$

For a value of p , the base suppression probability (z_i) ensures that consumers at higher risks would have more locations suppressed. The additional term ($z_i \times s_i^j$) ensures that a more informative location j is suppressed with a higher probability ($z_i + z_i \times s_i^j$). Since each consumer i 's risk r_i and the suppression weights \vec{s}_i can be computed apriori from the original unobfuscated data, the suppression probabilities above depend only on the grid parameter p . Suppressing the location data to limit a stalker's ability to invade private information would also adversely affect a butler advertiser's utility derived from $\mathcal{P}(T)$. For instance, in the extreme scenario when each consumer's risk $r_i = 1$ and p is reasonably high, all locations would be suppressed (i.e., complete suppression⁷: $\{\mathcal{P}(T_i)\} = \mathcal{P}(T) = \emptyset$), resulting in no utility to the advertiser, nor threat to consumer privacy. A similar inference can be made when $p = 0$ (i.e., no suppression: $\mathcal{P}(T) = T$), resulting in high data utility and also high privacy risk. Noting these two extreme scenarios, we empirically determine the suppression parameters $\{\vec{s}_i, z_i\}$ by varying the grid parameter p to derive the published trajectories $\mathcal{P}(T)$ that balance the risk and utility.

The proposed obfuscation scheme has two main advantages. First, the structured grid search by varying the grid parameter p provides the data collector with multiple trade-off choices. Second, the identified $\{\vec{s}_i, z_i\}$ provide the data collector with consumer level interpretability of the obfuscation. By

⁷Note that $s_i \in [0, 1]$; and $z_i \in [0, 1]$ because $r_i \in [0, 1]$ and $p \in [0, 1]$. Nonetheless, the corresponding location is suppressed with probability 1 whenever $(z_i + z_i \times s_i) > 1$.

fine-tuning $\{\vec{s}_i, z_i\}$, our ultimate goal is to understand, quantify, and optimize the trade-off between the data utility (\mathbf{U}) and privacy risk (\mathcal{PR}) in a meaningful way.

1.5 Empirical Study

Consistent with the proposed framework (Part A of Figure 1.1), prior to obfuscation, we first compute each consumer i 's baseline risk r_i (Section 1.4.1) and suppression weights \vec{s}_i (Section 1.4.3) on the unobfuscated data. We also compute the baseline data utility $MAP@k$ and $MAR@k$ across all consumers on the unobfuscated data (Section 1.4.2). Then for each $p \in_p = \{0, 0.1, \dots, 1\}$, we obfuscate each consumer's trajectory based on the suppression probabilities computed from the above r_i , \vec{s}_i and p (Equation 1.6); and re-compute the mean risk and utility across all consumers on the corresponding obfuscated data to assess the percentage decrease in the mean risk and utility from the baseline mean risk and baseline utility, respectively (Part B of Figure 1.1). This repeated process with varied p will offer the data collector multiple options to balance the risk and utility. We will report the details and key findings below.

1.5.1 Quantification of Consumer's Privacy Risk

As described above, for each type of threat, we quantify each consumer's baseline risk r_i and suppression weights \vec{s}_i without obfuscation. Then based on the suppression probabilities calculated from these r_i , \vec{s}_i , and each $p \in \{0, 0.1, \dots, 1\}$, we perform consumer-level obfuscation. Each p leads to a different set of obfuscated trajectories and hence re-computation of the mean risk and utility across consumers. To obtain consistent estimates of the mean risk and utility, we use bootstrapping with 20 trials for each p . In the sensitive attribute threat, we consider two sensitive attributes of home address and mobile operating system. To train the predictive model, we mimic a stalker with access to a training sample of known trajectories and sensitive attributes. We split the data into two random samples: 50% training set (T_{train}) with 20,000 consumers to train the predictive model, and 50% test set (T_{test}) with 20,012 consumers. As described in Section 1.4.1, we use Random Forest regressor to predict the risk of inferring home location and a Random Forest classifier to predict mobile operating system and the re-identification risk. We cross validate the models to avoid over-fitting by tuning model specific hyper parameters (see Appendix A.6 for more details). Once the model is trained, we apply it to estimate the risk on T_{test} in each privacy threat. In Figure 1.4, we report the average risk across all consumers in T_{test} for each p . To compute the re-identification risk, we assume the number of locations in each consumer's trajectory already known to a stalker is 2, that is, $|\bar{T}_i| = 2$ in Definition 2, to illustrate our approach.

A data collector can gain a host of insights from the initial step of quantifying consumers' privacy risks prior to obfuscation, such as which consumers are at the greatest risk, what is the severity of each privacy risk, which feature is most informative to a stalker and hence should be suppressed. For example, Figure 1.3a offers the data collector a visual of the distribution of

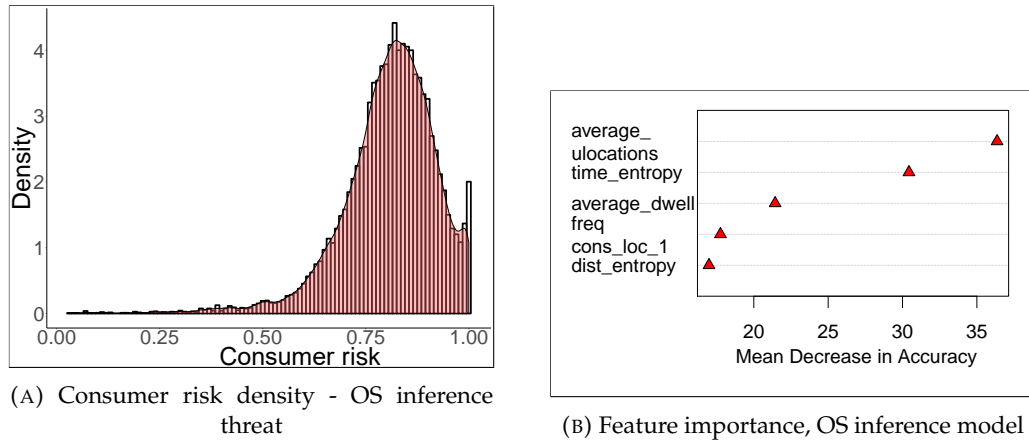


FIGURE 1.3: Personalized Risk Management Insights

the consumers' risks if a stalker were to infer their operating systems from the unobfuscated trajectory data. It shows that the majority of the consumers carry a relatively high risk (≥ 0.75 chance of success for a stalker) of their sensitive attribute of operating system being inferred if no obfuscation were performed. Also, the average risk of home address inference is 0.84. By assessing the error of the Random Forest regressor learned to predict the home address, we find that on average a stalker could successfully identify a consumer's home address within a radius of 2.5 miles (Appendix A.6). Further, the average risk of re-identifying an individual's entire trajectory by knowing merely two randomly sampled locations is 0.49, that is, a 49% chance of success for a stalker. In addition, the data collector can assess the worst cases associated with the top-risk consumers in each of the above threats.

Despite these paramount privacy risks arising from unobfuscated location data, they can be curtailed by a data collector using the proposed framework. For instance, the risk associated with the home address inference could be reduced by 10% while fully preserving the data utility on the POI@1 performance (Figures 1.4a, 1.4c, $p = 0.7$). As a follow-up step, by implementing the POI recommendation strategy in the real world, a data collector can also measure the monetary value of an individual trajectory, and compare it with the consumer-specific privacy risk to better understand the customer lifetime value (Berger and Nasr, 1998) and personalize customer relationship management.

In addition, a data collector may look at the feature importance (discussed in Appendix 1.4.1) prior to obfuscation. For instance, Figure 1.3b displays the top five most important features of the Random Forest trained to compute the consumers' risks. We observe that the top five features comprise of consumer mobility patterns and their affinity to various locations. Specifically, average number of unique locations visited by a consumer (`average_ulocations`), mobility entropy measuring the predictability of a consumer's trajectory (`time_entropy`, `dist_entropy`), average time spent at various locations (`average_dwell_freq`) as well as consumer affinity to different locations (`cons_loc_1`) are identified as important features in estimating the consumer risk for sensitive attribute inference. Based on these, a data collector can infer that the temporal information of the trajectories (`time_entropy` and `average_dwell`) contribute significantly to the model's predictive performance and consequently to the

consumer risk of identifying a sensitive attribute. Hence, a possible obfuscation scheme that removes (even partially) the timestamps in the trajectories would prevent the stalker from constructing the temporal features and potentially reduce the consumers' risks. Similar insights can be gained by analyzing the risk scores related to other stalker threats - home address inference and re-identification threat considered in the work.

1.5.2 Quantification of Advertiser's Utility

Next, we compute the data utility to a butler advertiser by leveraging a collaborative filtering recommendation heuristic, discussed in Section 1.5.2 to predict each consumer's future locations. To assess the predictive accuracy, we use the locations actually visited by each consumer in the fifth week as the ground truth and train the recommendation model to predict the locations. The model ranks the locations that a consumer is likely to visit in the fifth week of the observation period. We compute the average utility for the advertiser across all consumers, $MAP@k$ and $MAR@k$, for $k = \{1, 5, 10\}$ to illustrate the method's efficacy, where k is the next k locations. The model can also be used to compute $MAP@k$ and $MAR@k$ for other values of k . We perform 20 trials for each p and report the mean and 95% confidence intervals of the utility (Figure 1.4). A more detailed explanation of the utility computation is available in the Appendix A.5.

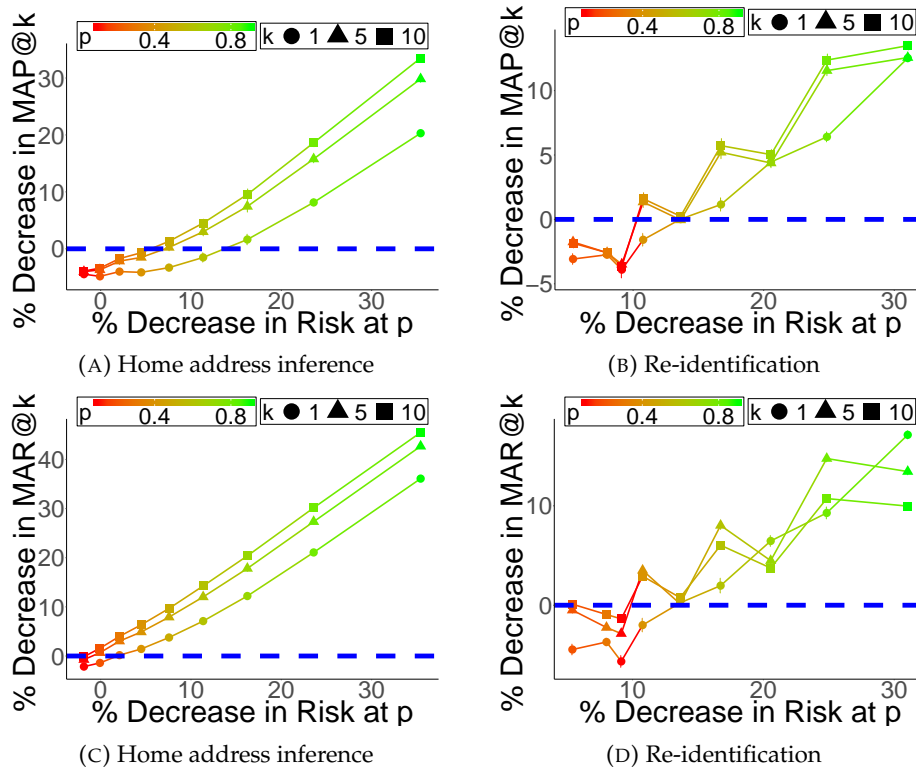


FIGURE 1.4: Proposed framework - $MAP@k$ and $MAR@k$ for varying p

1.5.3 Obfuscation Scheme for Data Collector

In Figures 1.4a and 1.4b, we visualize the risk-utility trade-off based on $MAP@k$. As described earlier, the locations in each T_i are suppressed based on the suppression probabilities computed from p and $\{\vec{s}_i, z_i\}$. We will focus on discussing the results where the suppression weights \vec{s}_i are computed based on the frequency to each location, although we have also computed \vec{s}_i based on recency and time spent at each location (Appendix 1.6.2). In Figures 1.4a and 1.4b, the X and Y axes display the percentage decrease in the mean risk and $MAP@k$ from the baseline risk and baseline $MAP@k$ for each $p \in p$. We plot these for $k = \{1, 5, 10\}$. Intuitively, the higher the value of X-axis, the more the decrease in the overall risk and hence better preservation of privacy. On the other hand, the lower values of Y-axis correspond to a lesser decrease in the utility of the obfuscated data compared to the original data, suggesting a similar utility for the advertiser even after obfuscation. A data collector who aims to trade off between utility and privacy is thus presented with multiple choices in our framework, with different k and p . Ideally, a good choice for obfuscation would be the values of p that correspond to a higher value along the X-axis and a lower value along the Y-axis. In the figures, the horizontal blue line, with no decrease in data utility from obfuscation indicates these choices. Similar insights can be drawn from figures 1.4c and 1.4d where we compare the percentage decreases in $MAR@k$ to the percentage decreases in the mean risk.

In all graphs in Figure 1.4, we observe that as we increase p , the values along both axes increase. This is expected since an increase in p , for the same consumer risk scores, more locations get suppressed, thus more information loss to an advertiser's utility as well as a privacy threat. For a given percentage decrease in risk, we observe a lesser corresponding percentage decrease in performance. This can be explained by the framework's obfuscation parameters $\{\vec{s}_i, z_i\}_{i=1}^N$ which are varied based on the consumer risk scores that capture the success of a privacy threat. This risk-based obfuscation would penalize and cause more information loss to the stalker's adversarial intent compared to the utility. The figures also emphasize the proposed framework's flexibility to provide a data collector with several interpretable choices for obfuscation. Further, since our obfuscation scheme works by suppressing a set of location tuples instead of randomization (Yang, Qu, and Cudre-Mauroux, 2018) or splitting (Terrovitis et al., 2017), this would also have potential benefits to the server costs incurred by an advertiser in storing and analyzing the location data.

1.5.4 Model Comparison

We compare the proposed obfuscation scheme with eight different baselines corresponding to three types of obfuscation approaches – obfuscation rules derived from timestamps of consumer locations, alternate suppression schemes based on consumer risk and the latest work in syntactic models LSUP and GSUP (Terrovitis et al., 2017). In each of the baselines, we are interested in 1) Percentage decrease in risk for the two types of consumer threats - home address risk and re-identification risk (estimated as discussed in Section 1.4.1) 2) Percentage decrease in advertiser's utility measured as $MAP@k$, $MAR@k$,

from their respective non-obfuscated consumer trajectories. For a specific privacy threat, a method is considered superior if the percentage decrease in advertiser’s utility is lesser compared to the percentage decrease in risks. We omit distortion, cloaking, synthesis based approaches discussed in Section 3.2 and limit our comparison to recent suppression based techniques since our advertiser’s utility is quantified based on POI prediction which involves prediction of a location rather than a coarser region.

Obfuscation rule	% Decrease Home address risk	% Decrease Re-identification risk	% Decrease Utility (MAP@1)	% Decrease Utility (MAR@1)
Remove Sleep hours	2.43	1.41	11.83	12.69
Remove Sleep and working hours	10.72	21.49	34.45	23.72
Remove time stamps	13.45	0	33.16	32.97

TABLE 1.4: Alternative Schemes: Rule-based Obfuscation

Comparison to Rule-based Obfuscations.

We derive a few practical rules for obfuscation based on the timestamps of the locations in the data. In the absence of a privacy-friendly framework, a data collector could perform obfuscation by choosing to 1) remove all the locations during the usual sleeping hours (10 PM - 7 AM) on all days, 2) remove the locations in sleeping hours and working hours (9 AM - 6 PM) on weekdays, or 3) remove the timestamps of the locations entirely before sharing the data. The three time-based rule obfuscations would reduce the amount of information that can be extracted from the shared location data, and hence adversely affect the advertiser’s utility. For instance, if the timestamps of the location data were to be removed, both the mobility features, `time_entropy`, `time_rrog`, `average_dwell` (from Table 1.3, Section 1.4.1) and the consumer-consumer, consumer-location affinity features based on time spent by a consumer at a location cannot be computed.

The decrease in risks for the two threats and decrease in utility for each of these obfuscations are presented in Table 1.4. As expected, there is a decrease in both the risk and utility. In the home address inference threat (Figure 1.4a, $p = 0.7$, $k = 1$), we find that a risk to consumer privacy can be reduced by 15% (maximum decrease when compared to the rule-based heuristics) with less than 1% decrease in $MAP@1$ (minimum decrease). A similar trend is observed in the re-identification threat (Figures 1.4b, 1.4d). Overall, we find a better choice set for the trade-off justifying a need for a privacy-friendly framework to assist a data collector to share location data in a privacy-friendly way.

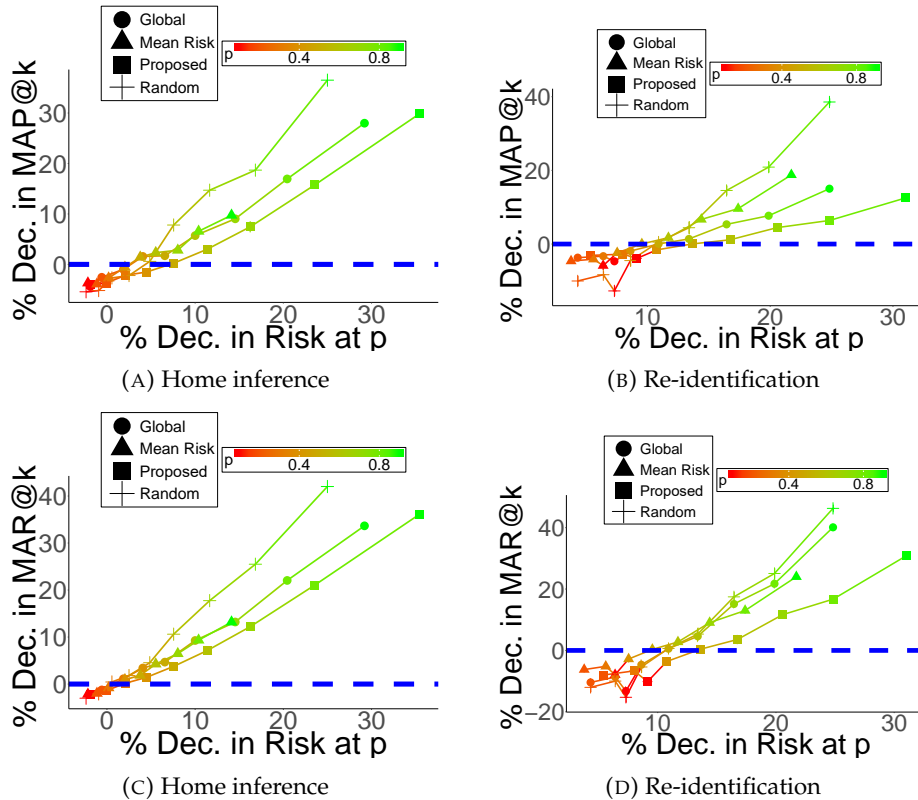


FIGURE 1.5: Proposed framework vs risk-based obfuscations - $MAP@1$ and $MAR@1$

Comparison to Risk-based Obfuscations.

We further compare the proposed obfuscation scheme to three other risk-based suppression baselines. These baselines are devised to show the efficacy of the quantification of each consumer's risk and personalized suppression achieved by introducing and identifying consumer-specific parameters $\{\vec{s}_i, z_i\}$ in our framework.

1. **Random** - In this baseline, we do not perform suppression of locations at a consumer level. Instead of hiding location tuples in T_i based on $z_i = r_i \times p$ and suppression weights \vec{s}_i , we randomly suppress locations in T . We suppress the same number of location tuples as in the proposed obfuscation scheme to make it comparable.
2. **Mean Risk** - Here, we perform a consumer-specific suppression without any variation across consumers. We replace the consumer risk score r_i with the mean $\bar{r} = \frac{1}{N} \sum_i r_i$ and suppress locations using $z = \bar{r} \times p$ and suppression weights \vec{s}_i for each T_i as described in Section 1.4.3.
3. **Global** - In this baseline, we suppress a location tuple globally. That is, a tuple in any T has the same chance of being suppressed irrespective of varied risk levels across consumers. This is different from the proposed obfuscation scheme where a tuple may not be suppressed for a lower risk consumer but has been suppressed for a higher risk consumer. For each tuple, we assign the mean of all consumers' risk scores as the tuple's risk score, vary p , and perform suppression.

We empirically compare the proposed obfuscation scheme to the baselines above and visualize $MAP@1$ and $MAR@1$ in Figure 1.5. We observe that, for a given decrease in risk, the proposed obfuscation has the least decrease in the utility gain across all three threats. Random baseline, which is an ablation of the proposed obfuscation scheme without the risk quantification step performs the worst among the alternative models. This justifies a need for threat quantification either at a consumer level (Mean Risk and proposed obfuscation) or at a location tuple level (Global). A performance better than the Mean Risk baseline shows that a personalized level of obfuscation for each consumer is necessary. Finally, a higher utility over the Global baseline emphasizes the need for quantifying and suppressing locations at a consumer level compared to a tuple level.

Comparison to Latest Suppression Models.

Next, we compare the proposed framework to the most recent syntactic models LSUP and GSUP proposed by Terrovitis et al., 2017. Both models obfuscate the location data to reduce the re-identification threat while maintaining utility. Methodologically, these models differ from the proposed framework (Section 1.4.3) in two ways. First, in both LSUP and GSUP, the consumer risk is only quantified for one threat (re-identification), whereas our framework additionally considers the sensitive attribute inference. Second, a location is suppressed either globally across all consumers (GSUP) or locally for a subset of consumers (LSUP). In contrast, our suppression scheme, due to the introduction of the two consumer specific parameters $\{\vec{s}_i, z_i\}$, suppression may be performed at a consumer level with varying suppression probabilities assigned to each location visited by the consumer. In addition, compared to the parsimonious input that our proposed framework requires, both models in consideration require multiple input parameters, such as the number of adversaries, background knowledge of each adversary in, and P_{br} that controls the number of locations suppressed either locally (LSUP) or globally (GSUP). The higher the value of P_{br} , the lower is the number of locations suppressed. In our comparison, we follow the empirical evaluation framework of the authors to set the number of adversaries, set the background knowledge of each adversary in, and merely vary P_{br} .

In Table 1.5⁸, we present the decrease in the consumer risk from the unobfuscated trajectories for the two types of privacy threats - re-identification and sensitive attribute inference⁹ and the corresponding measures of advertiser's utility as $MAP@1$, $MAR@1$. To identify the obfuscation scheme that provides the better/worse trade-off, we compute the slope ($\frac{Y}{X}$ in Figure 1.4 — % decrease in the utility divided by % decrease in the risk) for different decreases in the utility ($MAP@1$) of LSUP and GSUP. We observe that in a

⁸The authors in Terrovitis et al., 2017 consider four values of P_{br} in their work - {0.2, 0.25, 0.33, 0.50} and conclude that for a fixed number of adversaries, a higher data utility occurs at higher P_{br} values (less locations suppressed) while ensuring reduction in re-identification threat. The best value suggested in the paper was $P_{br} = 0.5$. Our choice of P_{br} was guided based on these experiments and observations.

⁹Since the considered models do not handle the sensitive attribute inference, we obfuscated the data to reduce the re-identification threat and use the same obfuscated data to quantify the reduction in the consumer risk for the two types of attacks.

Obfuscation Method	% Decrease Home address risk	% Decrease Re-identification risk	% Decrease Utility (MAP@1)	% Decrease Utility (MAR@1)
GSUP ($P_{br} = 0.2$)	18.12	14.52	7.74	8.31
GSUP ($P_{br} = 0.5$)	7.25	7.29	4.49	3.42
LSUP ($P_{br} = 0.2$)	22.16	31.56	5.31	7.12
LSUP ($P_{br} = 0.5$)	9.15	10.91	-1.65	0.86

TABLE 1.5: LSUP and GSUP comparison. (Green/Red indicate proposed framework provides a better/worse trade-off)

majority (6 out of 8) of the cases, the proposed framework provides a better trade-off (denoted by green color in Table 1.5) compared to both LSUP and GSUP. This improved trade-off comes with an added benefit that the proposed framework only requires one input parameter – the number of locations already known to a stalker in the re-identification threat, as compared to the great sets of parameters required by LSUP and GSUP.

1.6 Robustness Tests

1.6.1 Alternate Utility Function : Activity Prediction

We study the robustness of our proposed framework with another popular business application – Human activity prediction. Daily activities like commuting, working, eating, etc. can capture contextual information about a consumer rather than just location. Predicting these activities with a high accuracy can enable real-time, context-aware marketing campaigns. For example, if an advertiser can accurately predict when a consumer is likely going to a restaurant, based on their current GPS location, they could deliver a set of recommendations for the nearest restaurant chains with appropriate marketing messages. Similar to the POI recommendation, we imitate an advertiser’s use of location data for Human Activity prediction to quantify the utility. We first transform the location trajectories of each consumer T_i into activity trajectories A_i comprising of 14 different activity groups listed in Table 3.1. To map the locations to activities, we leverage Google Places API which fetches the place type of a location (second column in Table 3.1), which we then semantically group into the 14 activity groups (first column in Table 3.1), which capture a consumer’s day-day activities of consumption (restaurant, unhealthyactivities), leisure (recreation, personalcare, hotel, home, fitness), shopping (necessityshopping, leisureshopping) and commute (publictransport, owntransport) behavior.

We model the activity sequence of a consumer using a LSTM based architecture to jointly predict the following.

1. Next k activities of a consumer,

Activity group	Place type of location
hospital	hospital, doctor
health	physiotherapist, pharmacy, dentist, drugstore
necessityshopping	store, supermarket, convenience_store, home_goods_store, grocery_or_supermarket, hardware_store
fitness	gym
publictransport	transit_station, train_station, bus_station, light_rail_station, subway_station
owntransport	car_wash, car_repair, parking, gas_station, taxi_stand
religious	church, mosque, hindu_temple, synagogue
recreation	amusement_park, tourist_attraction, zoo, park, theatre, sports_stadium, concert, bowling_alley, art_gallery, aquarium, museum, movie_rental, book_store, library, movie_theater, campground
travel	hotel, lodging, rv
personalcare	beauty_salon, spa, hair_care
leisureshopping	clothing_store, department_store, shopping_mall, shoe_store, electronics_store, furniture_store
unhealthyactivities	casino, liquor_store, bar, night_club, cigarette
restaurant	restaurant, food, meal, bakery, cafe, meal_delivery, meal_takeaway
other	locations for which a place type was not identified by Places API

TABLE 1.6: Activity groups

2. Time of the day of each activity (morning, afternoon, evening),
3. Time of week of each activity (Weekend/Weekday)

Architecture

1. Input Layer: The LSTM model takes three input sequences that captures the state of a consumer, represented as a triplet (a, t, w) , where a , the activity is encoded as a one-hot vector of length 14, $t \in \{\text{Morning, Afternoon, Evening}\}$ the time of the day is encoded as 0/1 to indicate Weekend/Weekday. The input sequences are concatenated into a vector of length 18 and are fed into the next layer.
2. LSTM Layer: The encoded input then goes through an LSTM layer which has a hidden state to store historical information and is carried forward to subsequent time-steps.
3. Dropout Layer: The output of the LSTM layer is fed through a dropout layer to prevent the model from over-fitting on the training data by setting the activations of a certain percent of neurons (dropout rate) to zero.
4. Activated Dense Layer: The output of the Dropout layer is fed into a Dense layer which outputs a vector of length 18, representing the state triplet (a, t, w) . We apply a SoftMax activation over the first 14 elements and the next three elements separately - representing the probability assigned to each of the activities and time of the day of the activity. The last element has a rectified linear unit activation applied on it to represent the time of the week of the activity.

Training, Model Selection and Hyperparameter tuning

For each consumer, we randomly pick one week of their activities as their test set. Before training our models, a hold-out validation set, of 1-week worth of

activities per consumer is randomly separated out from the training dataset. To prevent overfitting, the models are trained until their performance on the validation set reaches a maximum. While training, we compute three losses – Categorical cross entropy for the activity, time of day and binary cross entropy for the time of the week. Since accuracy of activity prediction is more salient compared to the time of day and time of week, we assign disproportionate weights $(\lambda, 1-\lambda, 1-\lambda)$, $\lambda \in [0,1]$ while summing across these losses and empirically pick the λ with the best averaged validation accuracy (grid $\lambda \in [0.5, 0.6, 0.7]$). In addition to λ , we also tune the number of hidden states in the LSTM layer (grid: 64, 128, 256) and the drop-out rate (linear grid: 0.05 to 0.4, in increments of 0.05). The models were trained for a maximum of 500 epochs. The model with the best validation accuracy was then picked to be evaluated on the test set.

Utility Measurement

Similar to Section 1.4.1 (Eq. 4, 5), we compute the AP_i^k and AR_i^k for the three prediction tasks. The expected utility of all consumers' trajectories $E(u_i)$ is calculated as the average across the three prediction tasks, across all consumers, and is denoted as $MAP@k$ and $MAR@k$. These are assessed for varying p , k and their corresponding decreases in risks for both the sensitive attribute inference and re-identification threat in the figure 1.6.

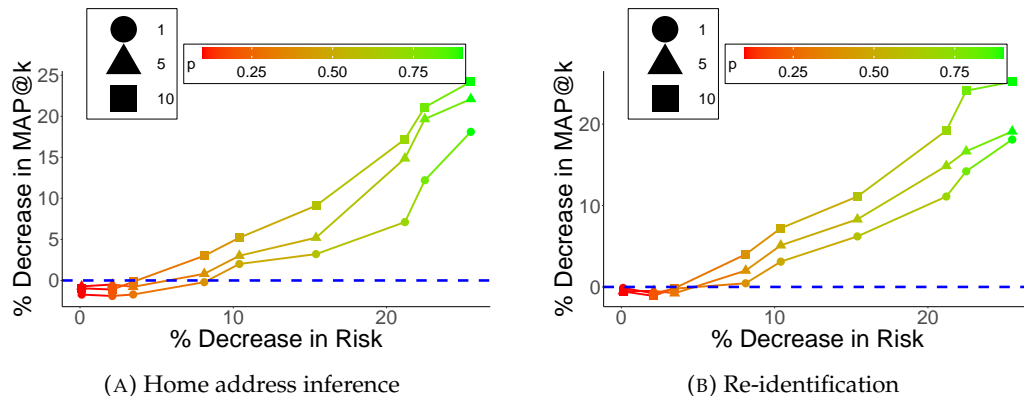


FIGURE 1.6: Proposed framework : Alternate Data Utility Function

Similar to Figure 1.4, for a given percentage decrease in risk, we observe a lesser corresponding percentage decrease in performance emphasizing the robustness of the proposed obfuscation scheme under time-aware recommendation heuristics.

1.6.2 Suppression based on Recency and Time Spent

In the proposed suppression scheme (Section 1.4.3), we introduce and provide a structured grid search by varying the grid parameter p to identify the two consumer specific parameters $\{\vec{s}_i, z_i\}$. Recall that z_i controls the number of locations to be suppressed for a given consumer trajectory T_i ; and within T_i , we assign weights to each tracked location through \vec{s}_i to denote the likelihood of a specific location being suppressed. In the Figure 1.4 of our empirical study (Section 1.5), we assign \vec{s}_i based on the frequency of the location

visited in T_i . Here, we further augment the empirical study and showcase the flexibility of the proposed suppression scheme by assigning the \bar{s}_i based on the time spent by a consumer at each location in T_i and the recency of the locations in T_i . For brevity, we only consider the sensitive attribute threat where a stalker aims to infer the home address of a consumer and visualize the privacy-utility trade-off in figures 1.7b and 1.7a. Similar to Figure 1.4, we observe that for a given percentage decrease in the risk, there is a lesser corresponding percentage decrease in the utility.

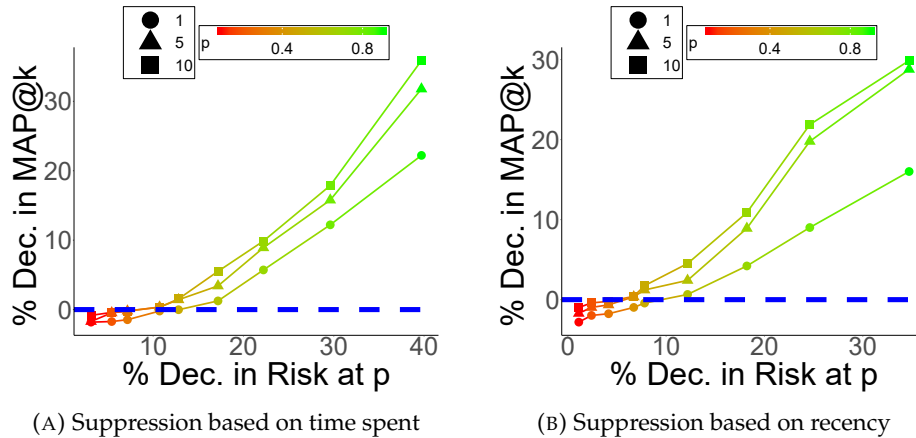


FIGURE 1.7: Proposed framework : Home address inference, suppression by recency and time spent.

1.6.3 Varying Sample Sizes

To test for the robustness of the results discussed in Figure 1.4, we repeat our empirical exercise on three random samples: 25%, 50% and 75% of the full 40,000 consumer trajectory data. For brevity and to avoid repetition of similar plots, the suppression is performed based on the frequency of the location visited by a consumer (similar to Figure 1.4) for the home address inference threat. The resulting plots comparing the percentage decreases in the consumer risk and advertiser utility from the baselines calculated on the unobfuscated data are visualized in Figures 1.8a, 1.8b, and 1.8c. We note that even at smaller samples, the slope (i.e., the % decrease in the utility divided by the % decrease in the risk) at different values of p is similar to that of the full sample (Figure 1.4a).

1.6.4 Varying Dimensionality

We exhibit the robustness of the proposed framework, by varying the dimensionality of the consumer trajectories. For each consumer trajectory, we perform 25%, 50% and 75% truncations and repeat our empirical exercise. For brevity, the suppression is performed based on the frequency of the location visited by a consumer (similar to Figure 1.4) for the home address inference threat. The resulting plots comparing the percentage decreases in the consumer risk and advertiser utility from the baselines calculated on the unobfuscated data are reported in Figures 1.9a, 1.9b, and 1.9c. We note that the proposed framework performs reasonably well on sparser dimensions

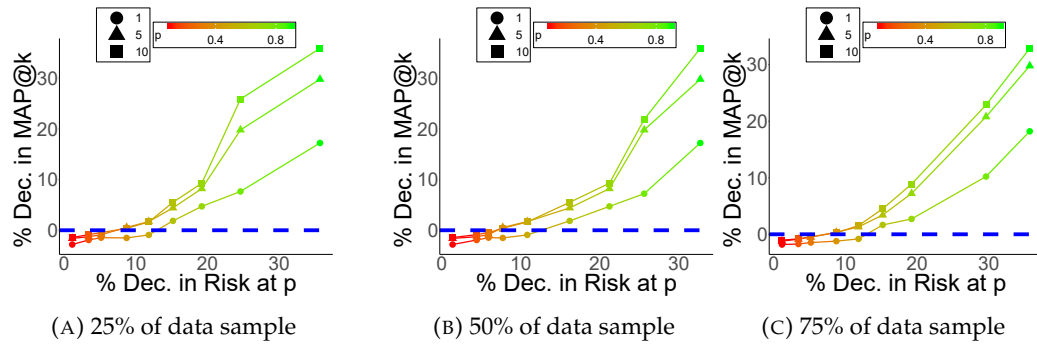


FIGURE 1.8: Proposed framework : Home address inference, varying sample sizes

(25% and 50%) and is comparable to the full sample (Figure 1.4a) when 75% of consumer trajectories are considered.

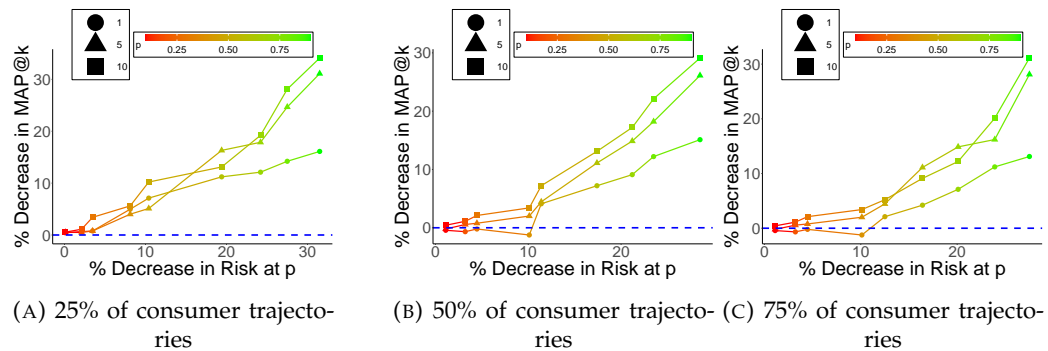


FIGURE 1.9: Proposed framework : Home address inference, varying sample sizes

1.7 Conclusion

Smartphone location tracking has created a wide range of opportunities for data collectors to monetize location data (Valentino-Devries et al., 2018; Thompson and Warzel, 2019). Leveraging the behavior-rich location data for targeting is proven to be an effective mobile marketing strategy and to increase advertisers' revenues (Ghose, Li, and Liu, 2018). However, these monetary gains come at the cost of potential invasion of consumer privacy. In this research, we tackle this important yet under-studied topic from a data collector's perspective. Specifically, we identify three key challenges facing a data collector and propose an end-to-end framework to enable a data collector to monetize and share location data with an advertiser while preserving consumer privacy.

The existing literature on privacy preservation is unsuited for this new type of data with distinct characteristics, not interpretable to the data collector, or not personalized to an individual level. Our research fills this gap. Specifically, we propose a framework of three components, each addressing a key challenge facing a data collector. First, we quantify each consumer's

risks, exemplified by two common types of stalker behaviors – sensitive attribute inference and re-identification threat. These risks are intuitively modeled as the stalker’s success probabilities in inferring the consumer’s private information. Second, we measure the utility of the location data to an advertiser by considering a popular business use case - POI recommendation. The utility is estimated by the accuracy of using the location data to infer a consumer’s future locations. Finally, to enable a data collector to trade off between consumer risk and advertiser utility, we propose an obfuscation scheme suppressing consumers’ trajectories based on their individual risk levels associated with each privacy threat and the informativeness of each location in their trajectories. The proposed obfuscation scheme also provides multiple options for the data collector to choose from based on specific business contexts.

We validate the proposed framework on a unique data set containing nearly a million mobile locations tracked from over 40,000 individuals over a period of five weeks in 2018. To our best knowledge, this research reflects an initial effort to analyze such a rich, granular, newly available human trajectory data, and for the purpose of privacy preservation. We find that there exists a high risk of invasion of privacy in the location data if a data collector does not obfuscate the data. On average, a stalker could accurately predict an individual’s home address within a radius of 2.5 miles and mobile operating system with an 82% success. The proposed risk quantification enables a data collector to identify high risk individuals and those features contributing most to the risk associated with each privacy threat. Furthermore, using the proposed obfuscation scheme, a data collector can achieve better trade-off between consumer privacy and advertiser utility when compared to several alternative rule-based and risk-based obfuscations. For instance, in the home address inference threat, we find that a risk to consumer privacy can be reduced by 15%, a maximum decrease when compared to rule-based heuristics, with less than 1% decrease in utility, a minimum decrease. Further, we compare the proposed framework with eight baselines and exemplify the performance gains in balancing the privacy-utility trade-off. In summary, this study presents conceptual, managerial, and methodological contributions to the literature and business practice, as summarized in the Introduction. Besides offering a powerful tool to data collectors to preserve consumer privacy while maintaining the usability of the increasingly accessible form of rich and highly valuable location data, this research also informs the ongoing debate of consumer privacy and data sharing regulations.

Despite the contributions, there are limitations of this research, thus calling for continued explorations of this rich and promising domain. For example, our data contain device IDs, but no detailed demographics, associated with each consumer. When such data become available, one may, for instance, develop deeper insights into which demographic sub-populations are most vulnerable to privacy risks. Also, our analysis considers the locations’ longitudes and latitudes, but not names (such as Starbucks) or types (such as hospital). Hence future research may further distinguish varied sensitivity levels across locations in privacy preservation. Furthermore, as other data, such as the same consumers’ online click streams or social media comments, become linked to their mobile location data, more sophisticated

privacy preservation methodologies may be developed. Lastly, in the proposed obfuscation scheme, the location data is obfuscated by assuming one consumer privacy threat and one advertiser objective at a time. That is, a composition of privacy threats or business objectives is not addressed. This calls for further methodological research to address this in the location data sharing paradigm.

Chapter 2

Explaining Anomalies in Groups

2.1 Introduction

Given a large dataset containing normal and labeled anomalous points, how can we *characterize* the anomalies? What combinations of features and feature values make the anomalies stand out? Are there *anomalous patterns*, that is, do anomalies form *groups*? How many different types of anomalies (or groups) are there, and how can we *describe* them succinctly for downstream investigation and decision-making by analysts?

Anomaly mining is important for numerous applications in security, medicine, finance, etc., for which many *detection* methods exist (Aggarwal, 2013). In this work, we consider a complementary problem to this vast body of work: the problem of anomaly *description*. Simply put, we aim to find human-interpretable explanations to already identified anomalies. Our goal is “reverse-engineering” known anomalies by unearthing their hidden characteristics—those that make them stand out.

The problem arises in a variety of scenarios, in which we obtain labeled anomalies, albeit no description of the anomalies that could facilitate their interpretation. Example scenarios are those where

- (1) the detection algorithm is a “black-box” and only provides labels, due to intellectual property or security reasons (e.g., Yelp’s review filter (Mukherjee et al., 2013)),
- (2) the detection algorithm does not produce an interpretable output and/or cannot explicitly identify anomalous patterns (e.g., ensemble detectors like bagged LOF (Lazarevic and Kumar, 2005) or isolation forest (Liu, Ting, and Zhou, 2008)), and
- (3) the anomalies are identified via external mechanisms (e.g., when software or compute-jobs on a cluster crash, loan customers default, credit card transactions get reported by card owners as fraudulent, products get reported by consumers as faulty, etc.). This setting also arises when security experts set up “honeypots” to attract malicious users, and later study their operating mechanisms (often manually). Examples include fake followers of honeypot Twitter accounts (Lee, Eoff, and Caverlee, 2011) and fraudulent bot-accounts that click honeypot ads (Dave, Guha, and Zhang, 2012).

Explaining anomalies is extremely useful in practice as anomalies are to be investigated by human analysts in almost all scenarios. Interpretation

of the anomalies help the analysts in sense-making and knowledge discovery, troubleshooting and decision making (e.g., planning and prioritizing actions), and building better prevention mechanisms (e.g., policy changes).

Our work taps into the gap between anomaly detection and its end usage by analysts, and introduces χ -PACS for *characterizing the anomalies in high-dimensional datasets*. Our emphasis is explaining the anomalies *in groups*¹. We model the anomalies to consist of (i) various patterns (i.e., sets of clustered anomalies) and (ii) outliers (i.e., scattered anomalies different from the rest). For example in fraud, malicious agents that follow similar strategies, or those who work together in “coalition”, exhibit similar properties and form anomalous groups. Bots deployed for e.g., click or email spam also tend to produce similar footprints as they follow the same source of command-and-control. At the same time, there may be multiple groups of fraudsters or bots with different strategies.

Explaining anomalies in groups has three key advantages: (1) it saves investigation time by providing a compact explanation, rather than the analyst having to go through anomalies one by one, (2) it provides insights into the characteristics of different anomaly types, and (3) importantly, it draws attention to anomalies that form patterns, which are potentially more critical as they are repetitive.

To lay out the challenges from a data mining perspective, we first introduce a list of desired properties (Desiderata 1–5) that approaches to the problem of anomaly explanation should satisfy. We then summarize our contributions.

2.1.1 Desiderata for Anomaly Description

In a nutshell, anomaly explanation methods should effectively characterize different kinds of anomalies present in the data, handle high dimensional datasets, and produce human-interpretable explanations that are distinct from normal patterns as well as succinct in length.

D1 Identifying different types of anomalies: Anomalies are generated by mechanisms other than the normal. Since such mechanisms can vary (e.g., different fraud schemes), it is likely for the anomalies to form multiple patterns in potentially different feature subspaces. A description algorithm should be able to characterize all types of anomalies.

D2 Handling high-dimensionality: Data instances typically have tens or even hundreds of features. It is meaningful to assume that the anomalies in a pattern exhibit only a (small) fraction of features in common. In other words, anomalies are likely to “hide” in sparse subspaces of the full space.

D3 Interpretable descriptions: It is critical that the explanation of the anomalies can be easily understood by analysts. In other words, descriptions should convey what makes a group of instances anomalous in a human-interpretable way.

D4 Discriminative (or detection) power: Explanations of anomalies should not also be valid for normal points. In other words, descriptions should

¹In this text, phrases ‘anomalous pattern’, ‘clustered anomalies’, and ‘group of anomalies’ are interchangeable.

be discriminative and separate the anomalies from the normal points sufficiently well. As a result, they could also help detect future anomalies of the same type.

D5 Succinct descriptions: It is particularly important to have simple and concise representations, for ease of visualization and avoiding information overload. This follows the Occam’s razor principle.

2.1.2 Limitations of Existing Techniques

Providing interpretable explanations for anomalies is a relatively new area of study compared to anomaly detection. However, the problem has similarities to description based techniques for imbalanced datasets. Almost all existing work in the anomaly detection literature assume anomalies to be scattered, and try to explain them one at a time (Dang et al., 2014; Dang et al., 2013; Keller et al., 2013; Knorr and Ng, 1999; Kuo and Davidson, 2016; Pevný and Kopp, 2014). Related work on collective data description (Görnitz, Kloft, and Brefeld, 2009; Tax and Duin, 2005), including rare class characterization (He and Carbonell, 2010; He, Tong, and Carbonell, 2010), assume a single pattern and/or do not look for subspaces. Other closely related areas are subgroup discovery and inductive rule learners. Subgroup discovery techniques (Gamberger and Lavrac, 2002; Herrera et al., 2011; Klösgen, 1996; Klösgen and May, 2002; Loekito and Bailey, 2008; Vreeken, Van Leeuwen, and Siebes, 2011) aim to describe individual classes while inductive rule learners (Cohen, 1995; Clark and Niblett, 1989; Deng, 2014; Friedman, Popescu, et al., 2008; Hara and Hayashi, 2016) focus on describing multiple classes with the aim of generalization rather than explanation. Another related line of work involves techniques for explaining black-box classifiers (Fong and Vedaldi, 2017; Koh and Liang, 2017; Montavon, Samek, and Müller, 2017; Ribeiro, Singh, and Guestrin, 2016) where the emphasis is on explaining the prediction of a single instance rather than a group of instances. Moreover, none of these work has an explicit emphasis on succinct, minimal descriptions.

To the best of our knowledge and as we expand in related work in §2.5, there is no existing work that provides a principled and general approach to the anomaly description problem that meet all of the goals in our desiderata adequately. (See Table 2.9 for an overview and comparison of related work.)

2.1.3 Summary of Contributions

Our work sets out to fill the gap, with the following main contributions:

- **Desiderata for Anomaly Description:** We introduce a new desiderata and target five rules-of-thumb (D1–D5) for designing our approach.
- **Description-in-Groups (DIG) Problem:** We formulate the explanation problem as one of identifying the various groups that the anomalies form (D1) within low-dimensional subspaces (D2).
- **Description Algorithm χ -PACS:** We introduce a new algorithm that produces interpretable rules (D3)—intervals on the features in each subspace, that are also discriminative (D4)—characterizing the anomalies in the group within a subspace but as few normal points as possible.

- **A New Encoding Scheme:** We design a new encoding-based objective for describing the anomalies in groups, based on the minimum description length (MDL) principle (Rissanen, 1978). Through non-monotone submodular optimization we carefully select the minimal subspace rules (D5) that require the fewest ‘bits’ to collectively describe all the anomalies.

Reproducibility: All of our code and datasets are open-sourced at <https://github.com/meghanathmacha/xPACS>.

2.2 Overview and Problem Statement

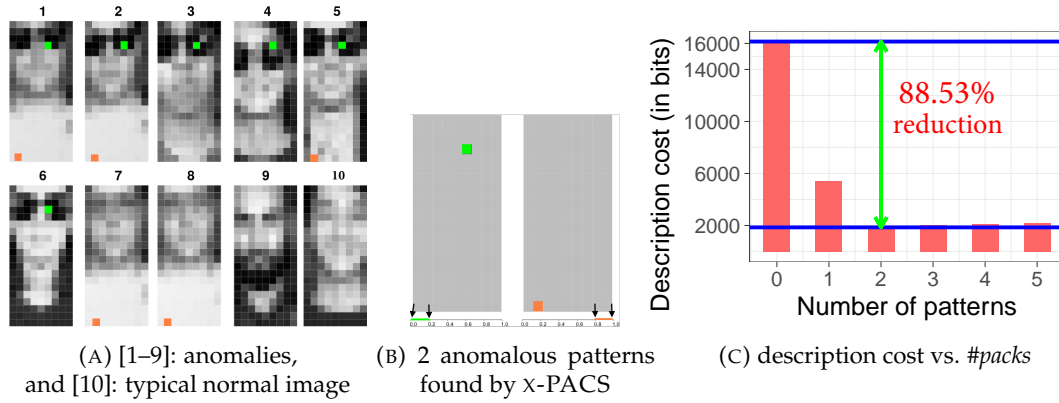
In a nutshell, *our goal is to identify a few, small micro-clusters of anomalies hidden in arbitrary feature subspaces that collectively and yet succinctly represent the anomalies and separate them from the normal points.* Specifically, our proposed χ -PACS finds a small set of low-dimensional hyper-ellipsoids (i.e., micro-clusters corresponding to *anomalous patterns* each enclosing a subset of the anomalies), and reveals scattered anomalies (i.e., outliers not contained in any ellipsoid).

Features that are part of the subspace in which a hyper-ellipsoid lies constitute its *characterizing subspace*. Ranges of values these features take are further characterized by the location (center and radii) of a hyper-ellipsoid within the subspace. Each hyper-ellipsoid is simply a “*pack of anomalies*” (hence the name χ -PACS²). The rest of the paper uses ‘hyper-ellipsoid’ or ‘*pack*’ in reference to anomalous patterns.

2.2.1 Example χ -PACS input-output

In Fig. 2.1 we show an example of the input and output of χ -PACS. We consider the face images dataset, in which χ -PACS identifies a minimum-description packing with two anomalous patterns and an outlier. In Fig. 2.1a, we visualize the dataset, where pixels are dimensions/features, that contains 9 labeled ‘anomalies’: [images 1–8] of 2 types (people w/ sunglasses or people w/ white t-shirt or both) + [image 9] an outlier (one person w/ beard). We also show [image 10], which is representative of 82 normal samples (people w/ black t-shirt w/out beard or sunglasses). In Fig. 2.1b, we display the anomalous patterns found by χ -PACS (characterizing subspaces are 1-d, feature rules/intervals shown at the bottom with arrows—the smaller, the darker the pixel) together explain anomalies 1–8 succinctly; 1-d pack (left) encloses images {1–6}, 1-d pack (right) encloses images {1,2,5,7,8}. Corresponding features/pixels highlighted on enclosed images. In Fig. 2.1c, we plot the description length (in bits, see §2.3.3) of anomalies individually (0 packs), vs. w/ 1–5 packs. χ -PACS automatically finds the best number of patterns (=2) that describe the anomalies and reveals the (unpacked) outlier [image 9].

² χ - refers to the number of packs, which we automatically identify via our data encoding scheme (§2.3.3). We use this naming convention after X-MEANS (Pelleg and Moore, 2000), which finds the number of k-means clusters automatically in an information-theoretic way.

FIGURE 2.1: Example χ -PACS input-output.(best in color)

2.2.2 Main Steps

Our χ -PACS consists of three main steps, each aiming to meet various criteria in our desiderata (D1–D5).

1. First we employ subspace clustering to automatically identify multiple clusters of anomalies (D1) embedded in various feature subspaces. Advantages of subspaces are two-fold: handling “curse of dimensionality” (D2) and explaining each pattern with only a few features (D5).
2. In the second step, we represent anomalies in each subspace cluster by an axis-aligned hyper-ellipsoid. Ellipsoids, in contrast to hyperballs, allow for varying spread of anomalies in each dimension. Axis-alignment ensures interpretable explanation with original features, which typically have real meaning to a user (D3). Moreover, we introduce a convex formulation to ensure that the ellipsoids are “pure” and enclose very few non-anomalous points, if at all, such that the characterization is discriminative (D4).
3. Final step is summarization, where we strive to generate minimal descriptions for ease of comprehension (D5). To decide which patterns describe the anomalies most succinctly, we introduce an encoding scheme based on the Minimum Description Length (MDL) principle (Rissanen, 1978). Our encoding-based objective lends itself to non-monotone submodular function maximization. Using an algorithm with approximation guarantees we identify a short list of patterns (hyper-ellipsoids) that are (i) compact with small radii, i.e., range of values that anomalies take per feature in the characterizing subspace is narrow; (ii) non-redundant, which “pack” (i.e., enclose) mostly different anomalies in various subspaces, and (iii) pure, which enclose either none or only a few normal points. Importantly, the necessary number of packs is automatically identified based on the MDL criterion.

Remark: Note that while χ -PACS identifies descriptive patterns of the anomalies, those can also be used for detection. Each pattern, along with its characterizing features and its enclosing boundary within that subspace can be seen as a discriminative *signature* (or set of rules), and can be used to label future instances—a new instance that falls within any of the *packs* is labeled

as anomalous. Instead of a single signature or an abstract classifier function or model, however, χ -PACS identifies multiple, interpretable signatures.

2.2.3 Notation and Definitions

Input dataset is denoted with $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$, containing m points in d -dimensions, where \mathcal{F} depicts the feature set. A subset $\mathcal{A} \subset \mathcal{D}$ of points are labeled as $y_{\mathcal{A}} = \text{'anomalous'}$, $|\mathcal{A}| = a$. The rest are $y_{\mathcal{D} \setminus \mathcal{A}} = \text{'normal'}$ points, denoted by \mathcal{N} , $|\mathcal{N}| = n$, $a + n = m$.

Our goal is to find “enclosing shapes”, called *packs*, that collectively contain as many of the anomalies as possible. While arbitrary shapes would allow for higher flexibility, we restrict these shapes to the hyper-ellipsoids family for ease of interpretation. This is not a strong limitation, however, since anomalous patterns are expected to form compact micro-clusters in some feature subspaces, rather than lie on arbitrarily shaped manifolds. A *pack* is formally defined as follows.

Definition 3 (pack) A pack p_k is a hyper-ellipsoid in a feature subspace $\mathcal{F}_k \subseteq \mathcal{F}$, $|\mathcal{F}_k| = d_k$, characterized by its center $\mathbf{c}_k \in \mathbb{R}^{d_k}$ and matrix $\mathbf{M}_k \in \mathbb{R}^{d_k \times d_k}$ where

$$p_k(\mathbf{c}_k, \mathbf{M}_k) = \{\mathbf{x} \mid (\mathbf{x} - \mathbf{c}_k)^T \mathbf{M}_k^{-1} (\mathbf{x} - \mathbf{c}_k) \leq 1\}.$$

We denote the anomalies that p_k encloses by $\mathcal{A}_k \subseteq \mathcal{A}$, and the normal points that it encloses by $\mathcal{N}_k \subset \mathcal{N}$.

Definition 4 (packing) A packing \mathcal{P} is a collection of packs as defined above; $\mathcal{P} = \{p_1(\mathbf{c}_1, \mathbf{M}_1), \dots, p_K(\mathbf{c}_K, \mathbf{M}_K)\}$ with size K .

2.2.4 Problem Statement

Based on the above definitions, our description-in-groups problem is formally:

DIG Problem 1 Given a dataset $\mathcal{D} \in \mathbb{R}^{m \times d}$ containing a anomalous points in \mathcal{A} and n non-anomalous or normal points in \mathcal{N} , $a \ll n$;

Find a set of anomalous patterns (packs) $\mathcal{P} = \{p_1, p_2, \dots, p_K\}$, each containing/enclosing a subset of the anomalies \mathcal{A}_k , where $\bigcup_{1 \leq k \leq K} \mathcal{A}_k \subseteq \mathcal{A}$,

such that \mathcal{P} provides the minimum description length $L(\mathcal{A} \mid \mathcal{D}, \mathcal{P})$ (in bits) for the anomalies in \mathcal{D} . (We introduce our MDL-based encoding scheme and cost function $L(\cdot)$ later in §2.3.3.)

Note that while packs enclose different subsets of anomalies in general, any two packs can have some anomalous points in common (since an anomaly can be explained in different ways), i.e., $\mathcal{A}_k \cap \mathcal{A}_l \neq \emptyset \ \exists k, l$. Packs can also share common features in their subspaces (as different types of anomalies may share some common characteristics), i.e., $\mathcal{F}_k \cap \mathcal{F}_l \neq \emptyset \ \exists k, l$. Moreover, the enclosing boundary of a *pack* may also contain some non-anomalous points. These issues related to the redundancy and purity of the packs would play a key role in the “description cost” of the anomalies. When it comes to identifying a small set of *packs* out of a list of candidates, we formulate

an encoding scheme as a guiding principle to selecting the smallest, least redundant, and the purest collection of *packs* that would yield the shortest description of all the anomalies.

2.3 χ -PACS: Explaining Anomalies in Groups

Next we present the details of χ -PACS, which consists of three building blocks:

§2.3.1 Subspace Clustering: Identify clusters of anomalies in various subspaces

§2.3.2 Refinement: Transform box-like subspace clusters to pure and compact hyper-ellipsoids (or *packs*)

§2.3.3 Summarization: Select subset of *packs* that yields the minimum description length of anomalies

We present our algorithms for each of these next.

2.3.1 Subspace Clustering: Finding Hyper-rectangles

In our formulation, we allow for anomalies to form multiple patterns, intuitively each containing anomalies of a different kind. We model anomalous patterns as compact “micro-clusters” in various feature subspaces.

In the first step, we use a subspace clustering algorithm, similar to CLIQUE (Agrawal et al., 1998) and ENCLUS (Cheng, Fu, and Zhang, 1999), that discovers subspaces with high-density anomaly clusters in a bottom-up, Apriori fashion. There are two main differences that we introduce. First, while prior techniques focus on a density (minimum count or mass) criterion, we use two criteria: (i) mass and (ii) purity, in order to find clusters that respectively contain many anomalous points, but also a low number of normal points. Second, we do not enforce a strict grid over the features but find varying-length high-density intervals through density estimation in a data-driven way.

Simply put, the search algorithm starts with identifying 1-dimensional intervals in each feature that meet a certain mass threshold. These intervals are then combined to generate 2-dimensional candidate rectangles. In general, k -dimensional hyper-rectangles are generated by merging $(k - 1)$ -dimensional ones that meet the mass criterion in a hierarchical fashion. Thanks to the monotonicity property of mass, the search space is pruned efficiently. Hyper-rectangles generated during the course of the bottom-up algorithm that meet *both* the mass and purity criteria are reported as clusters. A hyper-rectangle is formally defined as follows.

Definition 5 (hyper-rectangle) Let $\mathcal{F} = f_1 \times f_2 \times \dots \times f_d$ be our original d -dimensional numerical feature space. A hyper-rectangle $R = (s_1, s_2, \dots, s_{d'})$, $d' \leq d$, resides in a space $f_{t_1} \times f_{t_2} \times \dots \times f_{t_{d'}}$ where $t_i < t_j$ if $i < j$, and has d' sides, $s_z = [lb_z, ub_z]$, that correspond to individual intervals with lower and upper bounds in each dimension. A point $\mathbf{x} = \langle x_1, x_2, \dots, x_d \rangle$ is said to be contained or enclosed in hyper-rectangle $R = (s_1, s_2, \dots, s_{d'})$, if $lb_z \leq x_{t_z} \leq ub_z \quad \forall z = \{1, \dots, d'\}$.

Algorithm 1 SUBCLUS (\mathcal{D}, ms, μ)

Input: dataset $\mathcal{D} = \mathcal{A} \cup \mathcal{N} \in \mathbb{R}^{m \times d}$ with labeled anomalous and normal points, mass threshold $ms \in \mathbb{Z}$, purity threshold $\mu \in \mathbb{Z}$

Output: set of hyper-rectangles $\mathcal{R} = \{R_1, R_2, \dots\}$ each containing min. ms anomalous & max. μ normal points

- 1: Let $\mathcal{R}^{(k)}$ denote k -dimensional hyper-rectangles. Initialize $\mathcal{R}^{(1)}$ by kernel density estimation with varying quantile thresholds in $q = \{80, 85, 90, 95\}$, set $k = 1$
- 2: **for each** hyper-rectangle $R \in \mathcal{R}^{(k)}$ **do**
- 3: **if** mass(R) $\geq ms$ **then**
- 4: **if** impurity(R) $\leq \mu$ **then** $\mathcal{R}_{pure}^{(k)} = \mathcal{R}_{pure}^{(k)} \cup R$
 else $\mathcal{R}_{\neg pure}^{(k)} = \mathcal{R}_{\neg pure}^{(k)} \cup R$
- 5: **end if**
- 6: **end for**
- 7: $\mathcal{R} = \mathcal{R} \cup \mathcal{R}_{pure}^{(k)}$
- 8: $\mathcal{R}^{(k+1)} := \text{generateCandidates}(\mathcal{R}_{pure}^{(k)} \cup \mathcal{R}_{\neg pure}^{(k)})$
- 9: **if** $\mathcal{R}^{(k+1)} = \emptyset$ **then return** \mathcal{R}
- 10: $k = k + 1$, go to step 2

The outline of our subspace clustering is in Algorithm 1. It takes dataset \mathcal{D} as input with anomalous and normal points, a mass threshold ms equal to the minimum number of required anomalous points and a purity threshold μ equal to the maximum number of allowed normal points to be contained inside, and returns hyper-rectangles that meet the desired criteria.

To begin (line 1), we find 1-dimensional candidate hyper-rectangles, equivalent to intervals in individual features. To create promising candidate intervals initially, we find dense intervals with many anomalous points. To this end, we perform kernel density estimation (KDE³) on the anomalous points and extract the intervals of significant peaks.⁴ This is achieved by extracting the contiguous intervals in each dimension with density larger than the q -th quantile of all estimated densities. q is varied in $[80, 95]$ to obtain candidate intervals of varying length. An illustration is given in Fig. 2.2. Since multiple peaks may exist, multiple intervals can be generated per dimension as q is varied.

At any given level (or iteration) of the Apriori-like SUBCLUS algorithm, we scan all the candidates at that level (line 2–6) and filter out the ones that meet the mass criterion (line 3). Those that pass the filter are later merged to form candidates for the next level. Others with mass less than required are discarded, with no implications on accuracy. The correctness of the pruning procedure follows from the *downward closure property* of the mass criterion: for any k -dimensional hyper-rectangle with mass $\geq ms$, its projections in any one of $(k - 1)$ -dimensions must also have mass $\geq ms$.

At each level, we also keep track of the hyper-rectangles that meet both the mass and the purity criteria (line 4). Purity exhibits the *upward closure*

³KDE involves two parameters - the number of points sampled to construct the smooth curve and the kernel bandwidth. We set the sample size to 512 points and use the Silverman's rule of thumb (Silverman, 2018) to set the bandwidth.

⁴For categorical features, we would instead use histogram density estimation.

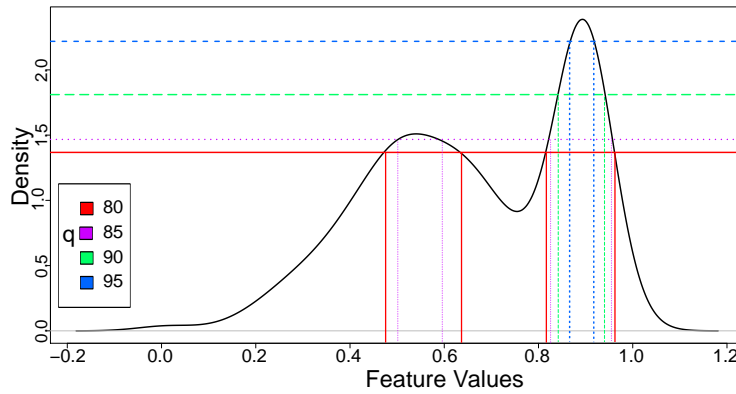


FIGURE 2.2: Identifying candidate hyper-rectangles in 1-d (equivalent to intervals) by KDE for varying quantile thresholds q .

property: for any $(k - 1)$ -dimensional hyper-rectangle that is pure (i.e. contains $\leq \mu$ normal points), any k -dimensional hyper-rectangle that subsumes it is also pure. This property could help us stop growing pure candidates by excluding them from the candidate generation step and speeding up the termination. While correct, however, such early-termination would prevent us from finding even purer hyper-rectangles later up in the hierarchy. To obtain as many candidate *packs* as possible, we continue our search for all hyper-rectangles that meet the mass criterion, and use the purity criterion for selecting the ones to be output (line 7).

The algorithm proceeds level by level. Having identified k -dimensional hyper-rectangles that satisfy the mass criterion, denoted $\mathcal{R}_{\geq ms}^{(k)} = \mathcal{R}_{pure}^{(k)} \cup \mathcal{R}_{-pure}^{(k)}$ (respectively for pure and not-pure sets), $(k + 1)$ -dimensional candidates are generated (line 8) in two steps: join and prune. The join step combines hyper-rectangles having first $(k - 1)$ dimensions as well as sides in common. That is, if $(s_{u_1}, s_{u_2}, \dots, s_{u_k})$ and $(s_{v_1}, s_{v_2}, \dots, s_{v_k})$ are two k -dimensional hyper-rectangles in $\mathcal{R}_{\geq ms}^{(k)}$, we require $u_i = v_i$ and $s_{u_i} = s_{v_i} \forall i \in \{1, \dots, (k - 1)\}$ and $u_k < v_k$ to form candidate $(k + 1)$ -dimensional hyper-rectangles of the form $(s_{u_1}, s_{u_2}, \dots, s_{u_k}, s_{v_k})$. The prune step discards all $(k + 1)$ -dimensional hyper-rectangles that have a k -dimensional projection outside $\mathcal{R}_{\geq ms}^{(k)}$. Again, the correctness of this procedure follows from the downward closure property of mass.

Choice of (ms, μ) : To obtain hyper-rectangles of varying size and quality, packing potentially different anomalies (and non-anomalies), we run Algorithm 1 with “conservative” parameters, i.e., low ms and high μ . As such, to generate a good volume of candidates, we set (ms, μ) as the median of the number of anomalous points, normal points from the 1-dimensional hyper rectangles. Setting a higher ms and lower μ would prune more (and potentially undesirably many) candidates in exchange of reduced time. We use the median to strike a balance between the quality and running time. As we describe later in §2.3.3, all these candidate packs are forwarded to a selection algorithm, which carefully chooses the subset that yields the shortest description of all the anomalies. As such, even though there are parameters input to Algorithm 1, we do not expect them from the user, rather we vary

and set those so as to generate various candidate packs. Having more candidates is likely to increase our chance of finding a combination that explains the anomalies the best (i.e., fewest bits).

To conclude, we note that other subgroup discovery techniques designed to handle numerical attributes, such as SubgroupMiner (Klösgen and May, 2002) or MIDOS (Wrobel, 1997), can be used as an alternative to Algo. 1; provided necessary modifications are incorporated to enforce the purity criterion.

2.3.2 Refining Hyper-rectangles into Hyper-ellipsoids

Grid or interval-based subspace clustering algorithms are limited to finding box-shaped rectangular clusters, and they may miss clusters inadequately oriented or shaped. To allow more flexibility, we refine each hyper-rectangle found by SUBCLUS into a hyper-ellipsoid (which we call a *pack*, recall Definition 3). An ellipsoid with center \mathbf{c} is written as

$$p(\mathbf{c}, \mathbf{M}) = \{\mathbf{x} \mid (\mathbf{x} - \mathbf{c})^T \mathbf{M}^{-1} (\mathbf{x} - \mathbf{c}) \leq 1\}$$

for positive semi-definite matrix $\mathbf{M} \succ 0$.

Given a hyper-rectangle R , let us denote the anomalous points it contains by $\mathbf{x}_i \in \mathcal{A}$ for $i = 1, \dots, a_R$ (See Def.n 5) and anomalous points outside R by $\mathbf{x}_j \in \mathcal{A}$ for $j = a_{R+1}, \dots, a$. The normal points are denoted by $\mathbf{x}_l \in \mathcal{N}$ for $l = 1, \dots, n$.

When we convert a given R to an ellipsoid, we would like all \mathbf{x}_i 's (anomalous points) it already contains to reside inside the ellipsoid. In contrast, we would like all \mathbf{x}_j 's (normal points) to remain outside the ellipsoid. The refinement is achieved by enclosing as many as the other anomalous points (\mathbf{x}_j 's) that are in the vicinity of R inside the ellipsoid as well. Those would be the points that were left out due to axis-aligned interval-based box shapes that hyper-rectangles are limited to capture. An illustration is given in Fig. 2.3.

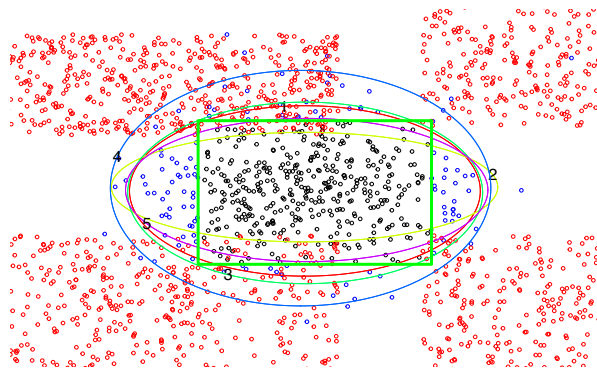


FIGURE 2.3: Example illustration of refining hyper-rectangles to ellipsoids in 2-d. Anomalous points (black) captured by SUBCLUS (Alg. 1) in a (green) rectangle, other anomalous points (blue) in the vicinity, and normal points (red).

First we describe our approach for \mathbf{x}_i 's and \mathbf{x}_l 's, the positive and negative points that we respectively aim to include and exclude. The goal is to find a discriminating function $h(\cdot)$ where $h(\mathbf{x}_i) > 0$ and $h(\mathbf{x}_l) < 0$. To this end, we use the quadratic function $h(\mathbf{x}) = \mathbf{x}^T \mathbf{U} \mathbf{x} + \mathbf{w}^T \mathbf{x} + w_0$, with parameters $\blacksquare = \{\mathbf{U}, \mathbf{w}, w_0\}$. We solve for \blacksquare by setting up an optimization problem based

on a semi-definite program (SDP), that satisfies $\mathbf{x}_i^T \mathbf{U} \mathbf{x}_i + \mathbf{w}^T \mathbf{x}_i + w_0 > 0$ for all i and $\mathbf{x}_l^T \mathbf{U} \mathbf{x}_l + \mathbf{w}^T \mathbf{x}_l + w_0 < 0$ for all l . Most SDP solvers do not work well with strict inequalities, thus we modify to a non-strict feasibility problem by adding a margin, and solve (for each R):

$$\begin{aligned} \min_{\mathbf{U}, \mathbf{w}, w_0, \varepsilon} \quad & \sum \varepsilon_i + \lambda \sum \varepsilon_l \\ \text{s.t.} \quad & \mathbf{x}_i^T \mathbf{U} \mathbf{x}_i + \mathbf{w}^T \mathbf{x}_i + w_0 \geq 1 - \varepsilon_i, \quad i = 1, \dots, a_R \\ & \mathbf{x}_l^T \mathbf{U} \mathbf{x}_l + \mathbf{w}^T \mathbf{x}_l + w_0 \leq -1 + \varepsilon_l, \quad l = 1, \dots, n \\ & \mathbf{U} \preceq -I, \quad \varepsilon_i \geq 0, \quad \varepsilon_l \geq 0 \end{aligned}$$

where \mathbf{U} is a negative semi-definite matrix. We can show that $(\mathbf{U}, \mathbf{w}, w_0)$ define an ellipsoidal enclosing boundary, wrapping \mathbf{x}_i 's inside and leaving \mathbf{x}_l 's outside, for which we allow some slack ε . λ is to account for the imbalance between the number of positive and negative samples. The optimization problem is convex, which we solve using an efficient off-the-shelf solver, where each hyper-rectangle output by SUBCLUS can be processed independently.

Having set up our refinement step as a convex quadratic discrimination problem, we next describe how we incorporate \mathbf{x}_j 's (anomalous points outside R) into the optimization. Intuitively, we would like to include as many other anomalies as possible inside the ellipsoid, but only those that are nearby \mathbf{x}_i 's and not necessarily those that are far away. In other words, we only want to "recover" the \mathbf{x}_j 's surrounding a given R and not grow the ellipsoid to include far away \mathbf{x}_j 's to the extent that it would end up including many normal points as well.

To this end, we treat \mathbf{x}_j 's similar to \mathbf{x}_i 's but incur a lower penalty of excluding an \mathbf{x}_j than excluding an \mathbf{x}_i or including an \mathbf{x}_l . The optimization is re-written as

$$\begin{aligned} \min_{\mathbf{U}, \mathbf{w}, w_0, \varepsilon} \quad & \sum \varepsilon_i + \alpha \sum \varepsilon_j + \lambda \sum \varepsilon_l \\ \text{s.t.} \quad & \mathbf{x}_i^T \mathbf{U} \mathbf{x}_i + \mathbf{w}^T \mathbf{x}_i + w_0 \geq 1 - \varepsilon_i, \quad i = 1, \dots, a_R \\ & \mathbf{x}_j^T \mathbf{U} \mathbf{x}_j + \mathbf{w}^T \mathbf{x}_j + w_0 \geq 1 - \varepsilon_j, \quad j = a_{R+1}, \dots, a \\ & \mathbf{x}_l^T \mathbf{U} \mathbf{x}_l + \mathbf{w}^T \mathbf{x}_l + w_0 \leq -1 + \varepsilon_l, \quad l = 1, \dots, n \\ & \mathbf{U} \preceq -I, \quad \varepsilon_i \geq 0, \quad \varepsilon_j \geq 0, \quad \varepsilon_l \geq 0 \end{aligned}$$

Here, setting α (penalty constant for \mathbf{x}_j 's) smaller than both 1 and λ is likely a good choice. However, we do not know which (α, λ) pair would provide a good trade-off in general. Therefore, we sweep over a grid of possible values⁵ and generate various ellipsoids, as illustrated for the example case in Fig. 2.3. A last but important step is to sweep over the collection to discard *dominated* packs. Specifically, we output only the set of p 's in the Pareto frontier w.r.t. mass versus purity. In this set there are *no two packs where one*

⁵We use $\alpha = \{10^{-6}, 10^{-5}, \dots, 1\} \times \lambda = \{10^{-3}, 10^{-2}, \dots, 10^3\}$.

strictly dominates the other—by enclosing both higher number of anomalous points (higher mass) and lower number of normal points (higher purity).

We refine a hyper-rectangle $R = (s_1, s_2, \dots, s_{d'})$ into an ellipsoid within the same subspace, in other words, $\mathbf{U} \in \mathbb{R}^{d' \times d'}$ and $\mathbf{w} \in \mathbb{R}^{d'}$. For interpretability, we constrain \mathbf{U} to be diagonal to obtain *axis-aligned* ellipsoids as shown in Fig. 2.3, since the original features have meaning to the user.⁶

Our explanation consists of one rule on each feature in the subspace. A feature rule is a \pm radius interval around the ellipsoid's center. Formally:

Definition 6 (Feature rules) *Given an axis-aligned ellipsoid $p(\mathbf{c}, \mathbf{M})$ in a subspace $f_{t_1} \times \dots \times f_{t_{d'}}$, a rule on feature t_z is an interval $(\mathbf{c}[z] - \text{radius}_z, \mathbf{c}[z] + \text{radius}_z)$, where $\text{radius}_z = \sqrt{\mathbf{M}_{zz}}$, $\forall z = \{1, \dots, d'\}$. Conjunction of all d' feature rules constitute the signature of p .*

To wrap up, we show how to compute \mathbf{c} and \mathbf{M}^{-1} from $(\mathbf{U}, \mathbf{w}, w_0)$ to obtain the center and radii for an ellipsoid, using which we generate the feature rules.

Obtaining \mathbf{c} : At the boundary of the ellipsoid, $h(\mathbf{x}) = 0$ and inside $h(\mathbf{x}) > 0$. Center is the point where $h(\mathbf{x})$ is the maximum. Hence;

$$\mathbf{c} := \max_{\mathbf{x}} \mathbf{x}^T \mathbf{U} \mathbf{x} + \mathbf{w}^T \mathbf{x} + w_0 = -\frac{1}{2} \mathbf{U}^{-1} \mathbf{w} \quad (2.1)$$

Obtaining \mathbf{M}^{-1} :

$$\begin{aligned} \mathbf{x}^T (-\mathbf{U}) \mathbf{x} - \mathbf{w}^T \mathbf{x} - w_0 &< 0 & (2.2) \\ \mathbf{x}^T (-\mathbf{U}) \mathbf{x} + 2\mathbf{c}^T \mathbf{U} \mathbf{x} - w_0 &< 0 \quad \text{using Eq. (2.1)} \\ (\mathbf{x} - \mathbf{c})^T (-\mathbf{U}) (\mathbf{x} - \mathbf{c}) + \mathbf{c}^T \mathbf{U} \mathbf{c} - w_0 &< 0 \\ (\mathbf{x} - \mathbf{c})^T \frac{-\mathbf{U}}{(w_0 - \mathbf{c}^T \mathbf{U} \mathbf{c})} (\mathbf{x} - \mathbf{c}) &< 1 \implies \mathbf{M}^{-1} = \frac{-\mathbf{U}}{(w_0 - \mathbf{c}^T \mathbf{U} \mathbf{c})} \end{aligned}$$

Obtaining radii:

$$(\mathbf{x} - \mathbf{c})^T \mathbf{M}^{-1} (\mathbf{x} - \mathbf{c}) = \sum_{z=1}^{d'} (\mathbf{x}[z] - \mathbf{c}[z])^2 (\mathbf{M}^{-1})_{zz} \leq 1$$

To compute radius in dimension z , we find point \mathbf{x} where $\mathbf{x}[z'] = \mathbf{c}[z']$, $\forall z' \neq z$, and $(\mathbf{x}[z] - \mathbf{c}[z])^2 \frac{1}{\mathbf{M}_{zz}} = 1$. It is easy to see that $\text{radius}_z = |\mathbf{x}[z] - \mathbf{c}[z]| = \sqrt{\mathbf{M}_{zz}}$.

2.3.3 Summarization: Pack Selection for Shortest Description

Our ultimate goal is to find anomalous patterns that explain or summarize the given anomalies in the dataset as succinctly as possible. Intuitively, “good” patterns enclose similar groups of points and hence help compress the data. To this end, we formulate our summarization objective by an encoding scheme

⁶If the anomalous patterns are to be used for detection, we estimate a full \mathbf{U} matrix (i.e., possibly rotated ellipsoid).

and then devise an algorithm that carefully chooses a few patterns, in particular *packs* produced in §2.3.2, that yield the minimum encoding length. In the following, we describe our encoding scheme, followed by the proposed subset selection algorithm.

MDL formulation for encoding a given *packing*

Our encoding scheme involves a Sender (us) and a Receiver (remote). We assume both of them have access to dataset $\mathcal{D} \in \mathbb{R}^{m \times d}$ but only the Sender knows the set of anomalous points \mathcal{A} . The goal of the Sender is to transmit (over a channel) to the Receiver the information about which points are the anomalies *using as few bits as possible*. Naïvely encoding all feature values of every anomalous point *individually* would cost $|\mathcal{A}|d \log_2 f$ bits.⁷ The idea is that by encoding the enclosing boundary of *packs* (ellipsoids) found in §2.3.2, we (the Sender) could have the Receiver identify the anomalies *in groups*, which could save bits.

Obviously we would want to avoid “noisy” *packs* that include many normal points—that would necessitate spending extra bits for encoding those exceptions (i.e. “telling” the Receiver which points in a pack are *not* anomalies). Moreover, we would want to avoid using *packs* that encode largely overlapping group of anomalies, as bits would be wasted to redundancy. While identifying the *packing* that yields the fewest bits is the main problem, we first lay out our description length objective, for a given *packing* $\mathcal{P} = \{p_1(\mathbf{c}_1, \mathbf{M}_1), \dots, p_K(\mathbf{c}_K, \mathbf{M}_K)\}$:

- Transmit number of *packs* = $\log^* K$ ⁸
- For each pack $p_k \in \mathcal{P}$:
 - Transmit number of dimensions = $\log^* d_k, d_k \leq d$
 - Transmit *identity* of dimensions = $\log_2 \binom{d}{d_k}$
 - Transmit the center $\mathbf{c}_k = d_k \log_2 f$
 - Transmit $\mathbf{M}_k = d_k^2 \log_2 f$ ($d_k \log_2 f$ if diagonal)
 - Transmit exceptions (i.e., non-anomalies in p_k):
 - * number of normal points in $p_k = \log^* n_k$
 - * *identity* of normal points; by forming all possible subsets of size n_k of m_k (total number of points in p_k) = $\log_2 \binom{m_k}{n_k}$ (based on a canonical ordering of subsets, where points are ordered by distance to center)⁹

⁷Value of f is chosen according to the required floating point precision in the normalized feature space \mathbb{R}^d .

⁸Cost of encoding an arbitrary integer K is $L_{\mathbb{N}}(K) = \log^*(K) + \log_2(c)$, where $c \approx 2.865064$ and $\log^*(K) = \log_2(K) + \log_2(\log_2(K)) + \dots$ summing only the positive terms (Rissanen, 1978). We drop $\log_2(c)$ as it is constant for all *packings*.

⁹Another way to identify the normal points in a *pack*: sort points by their distance to center and send the index of normal points in this list of length m_k . This costs more for $n_k \geq 2$: $n_k \log_2 m_k > \log_2 \frac{m_k^{n_k}}{n_k!} > \log_2 \binom{m_k}{n_k}$.

Total cost of encoding with *packing* \mathcal{P} is then

$$\ell(\mathcal{P}) = \log^* K + \sum_{k=1}^K L(p_k), \quad \text{where} \quad (2.3)$$

$$L(p_k) = \log^* d_k + \log_2 \binom{d}{d_k} + d_k(d_k + 1) \log_2 f + \log^* n_k + \log_2 \binom{m_k}{n_k} \quad (2.4)$$

MDL objective function

Our objective is to find a *packing*, that is to identify a subset of *packs*, which provides the minimum encoding length. However, we do not assume that all anomalies would be covered by a packing, i.e., $\bigcup_k \mathcal{A}_k \subseteq \mathcal{A}$, as there could be anomalous points (outliers) that do not belong in any pattern but lie away from the others. The outliers $\mathcal{A} \setminus \{\bigcup_k \mathcal{A}_k\}$ are yet to be encoded individually. Description length of all anomalies \mathcal{A} with *packing* \mathcal{P} is

$$L(\mathcal{A}|\mathcal{D}, \mathcal{P}) = (|\mathcal{A}| - |\bigcup_{p \in \mathcal{P}} \mathcal{A}_p|) d \log_2 f + [\log^* |\mathcal{P}| + \sum_{p \in \mathcal{P}} L(p)]$$

where the second term [in brackets] is $\ell(\mathcal{P})$: cost of transmitting \mathcal{P} (and the anomalies covered by it) by Eq. (2.3), and the first term is the cost of individually encoding the remaining anomalies not covered by \mathcal{P} .

Notice that the objective of finding a subset \mathcal{S} that minimizes the description length is equivalent to selecting a *packing* that reduces the naïve encoding cost of $|\mathcal{A}| d \log_2 f$ the most, i.e.:

$$\boxed{\max_{\mathcal{S}} R_{\ell}(\mathcal{S}) = |\bigcup_{p \in \mathcal{S}} \mathcal{A}_p| c_u - \log^* |\mathcal{S}| - \sum_{p \in \mathcal{S}} L(p) + [\log^* |\mathcal{E}| + \sum_{p' \in \mathcal{E}} L(p')]} \quad (2.5)$$

where $c_u = d \log_2 f$ is a constant unit-cost to encode a point, and set \mathcal{E} denotes all the ellipsoids returned from the second part (refinement), as such, $\mathcal{S} \subseteq \mathcal{E}$. First three terms of the objective capture the overall *reduction* in encoding cost due to the packing with ellipsoids in \mathcal{S} . We can read it as aiming to *find a packing that covers as many anomalies as possible (expressive), while having small model cost (low complexity)—containing only a few packs in low dimensions*. The constant term [in brackets] ensures that $R_{\ell}(\mathcal{S})$ is a non-negative function.

Subset selection algorithm

for MDL *packing*

To devise a subset selection algorithm, we start by studying the properties of our objective function R_{ℓ} , such as submodularity and monotonicity that could enable us to use fast heuristics with approximation guarantees. Unfortunately, R_{ℓ} is not submodular as it is given in Eq. (2.5). However, with a slight modification where we fix the solution size (number of output

packs) to $|\mathcal{S}| = K$, such that the second term is constant $\log^* K$, the function becomes submodular, as we show below.

Theorem 1 *Our cardinality-constrained objective set function $R'_\ell(\mathcal{S})$ is submodular. That is, for all subsets $\mathcal{S} \subseteq \mathcal{T} \subseteq \mathcal{E}$ and packs $p \in \mathcal{E} \setminus \mathcal{T}$, it holds that*

$$R'_\ell(\mathcal{S} \cup \{p\}) - R'_\ell(\mathcal{S}) \geq R'_\ell(\mathcal{T} \cup \{p\}) - R'_\ell(\mathcal{T}).$$

Proof 1 Let $\text{Cover}(\mathcal{S}) = |\bigcup_{p \in \mathcal{S}} \mathcal{A}_p|$ return the number of anomalies contained by the union of packs in \mathcal{S} . Canceling the equivalent terms and constants on each side of the inequality, we are left with $\text{Cover}(\mathcal{S} \cup \{p\}) - \text{Cover}(\mathcal{S}) \geq \text{Cover}(\mathcal{T} \cup \{p\}) - \text{Cover}(\mathcal{T})$. The inequality follows from the submodularity property of the Cover function. ■

It is also easy to see that R'_ℓ is not monotonic.

Theorem 2 *Our modified objective set function $R'_\ell(\mathcal{S})$ is non-monotonic. That is, there exists $\exists \mathcal{S} \subseteq \mathcal{T}$ where $R'_\ell(\mathcal{T}) < R'_\ell(\mathcal{S})$.*

Proof 2 For $\mathcal{S} \subseteq \mathcal{T}$, $\text{Cover}(\mathcal{T}) \geq \text{Cover}(\mathcal{S})$ due to monotonicity of Cover function. On the other hand, description cost of packs in \mathcal{T} is $\sum_{p \in \mathcal{T}} L(p) = \sum_{p' \in \mathcal{S}} L(p') + \sum_{p'' \in \mathcal{T} \setminus \mathcal{S}} L(p'')$ and hence is strictly greater than those of \mathcal{S} . As such, for two packings $\mathcal{S} \subset \mathcal{T}$ with the same coverage, we would have $R'_\ell(\mathcal{T}) < R'_\ell(\mathcal{S})$.¹⁰ ■

Maximizing a submodular function is NP-hard as it captures problems such as Max-Cut and Max k-cover (Gharan and Vondrak, 2011). Nevertheless the structure of submodular functions makes it possible to achieve non-trivial results. In particular, there exist approximation algorithms for *non-monotone submodular* functions that are *non-negative*, like our objective function R'_ℓ . In particular, one can achieve an approximation factor of 0.41 for the maximization of any non-negative non-monotone submodular function *without* constraints (Gharan and Vondrak, 2011).

In our case, we need to solve our objective under the cardinality (i.e., subset size) constraint, where $|\mathcal{S}|$ is fixed to some K (since only then R_ℓ is submodular). To this end, we use the RANDOM-GREEDY algorithm by Buchbinder et al. (Buchbinder et al., 2014), which provides the best known guarantee for the cardinality-constrained setting, with approximation factors in $[0.356, \frac{1}{2} - o(1)]$. The algorithm is quite simple; at each step of K iterations, it computes the marginal gain of adding a single *pack* $p \in \mathcal{E} \setminus \mathcal{S}$ to \mathcal{S} and selects one among the top K highest-gain *packs* uniformly at random.

Choice of K : We identify K , the number of *packs* to describe the anomalies, automatically, best of which is unknown apriori. Concretely, we solve to obtain subset \mathcal{S}_K^* each time for a fixed $K = |\mathcal{S}_K^*| = 1, 2, \dots, a$, and return the solution with the largest objective value of $R_\ell(\mathcal{S}_K^*) = R'_\ell(\mathcal{S}_K^*) - \log^* K$ in Eq. (2.5). This is analogous to model selection with regularization for increasing model size.

¹⁰Intuitively, this is where R_ℓ drops when we add a new *pack* to \mathcal{S} (with positive cost) that does not cover any new anomalies.

2.3.4 Overall Algorithm χ -PACS

Algorithm 2 puts together all three components of χ -PACS as described through §2.3.1–§2.3.3. We conclude this section with the computational complexity analysis.

Algorithm 2 χ -PACS ($\mathcal{A} \cup \mathcal{N}$): Explaining Anomalous Patterns

Input: dataset $\mathcal{D} = \mathcal{A} \cup \mathcal{N}$ with labeled anomalies

Output: set of anomalous patterns (represented as hyper-ellipsoids)

$\mathcal{P} = \{p_1(\mathbf{c}_1, \mathbf{M}_1), \dots, p_K(\mathbf{c}_K, \mathbf{M}_K)\}$

- 1: Set of hyper-rectangles $\mathcal{R} = \emptyset$
- 2: Obtain $\mathcal{R}^{(1)}$ (1-d intervals) by kernel density estimation, varying cut-off threshold in $q = \{80, 85, 90, 95\}$
- 3: $\hat{f}_a :=$ distribution of number of anomalies across $\mathcal{R}^{(1)}$
- 4: $\hat{f}_n :=$ distribution of number of normal points across $\mathcal{R}^{(1)}$
- 5: $\mathcal{R} := \text{SUBCLUS}(\mathcal{D}, ms = q(\hat{f}_a, 50), \mu = q(\hat{f}_n, 50))$ by Alg. 1 in §2.3.1
- 6: Set of hyper-ellipsoids $\mathcal{E} = \emptyset$
- 7: **for** $R \in \mathcal{R}$ **do**
- 8: $\mathcal{E}_R = \emptyset$
- 9: **for** $\alpha = \{10^{-6}, 10^{-5}, \dots, 1\}$ **do**
- 10: **for** $\lambda = \{10^{-3}, 10^{-2}, \dots, 10^3\}$ **do**
- 11: $\mathcal{E}_R := \mathcal{E}_R \cup$ solve optimization problem in §2.3.2 for (R, α, λ)
- 12: **end for**
- 13: **end for**
- 14: $\mathcal{E} := \mathcal{E} \cup \text{ParetoFrontier}(\mathcal{E}_R)$
- 15: **end for**
- 16: **for** $K = 1, \dots, |\mathcal{A}|$: select a subset $\mathcal{S}_K^* \subset \mathcal{E}$ of K packs using the cardinality-constrained RANDOM-GREEDY algorithm by Buchbinder et al. (Buchbinder et al., 2014) to optimize the description length reduction objective $R_\ell(\cdot)$ in §2.3.3.
- 17: **return** $\mathcal{P} := \arg \max_{\mathcal{S}_K^*} R'_\ell(\mathcal{S}_K^*) - \log^* K$

Computational complexity: We analyze the complexity of each part separately. Main computation of §2.3.1 is the SUBCLUS algorithm. Preliminary KDE to create 1-d intervals is independently done per dimension in parallel, only on the anomalous points. We use a constant number of sampling locations, as such, KDE complexity is $O(a)$ where a is the number of anomalies. SUBCLUS then proceeds level-by-level and makes as many passes over the data as the number of levels. For a d' dimensional hyper-rectangle that meets the mass and purity criteria, all its $2^{d'}$ projections in any subset of the dimensions also meet the mass criterion (although may not be pure). As such, running time of SUBCLUS is exponential in the highest dimensionality of the hyper-rectangle that meets both criteria. Total time complexity of this step is $O(c^{d_{\max}} + md_{\max})$ for a constant c ¹¹ that accounts for possibly multiple d_{\max} -dimensional hyper-rectangles and the smaller ones. The second term captures the passes over the data over d_{\max} levels.

The main computation of §2.3.2 is solving the SDP optimization problem, for which we use the popular cvx SDPT3 solver that takes $O((d_{\max} + m)^3)$

¹¹For instance, if we have t d_{\max} -dimensional hyper-rectangles, then the complexity would be $O(t2^{d_{\max}} + md_{\max})$, we could rewrite this as $O(c^{d_{\max}} + md_{\max})$

for an axis-aligned ellipsoid (or diagonal \mathbf{U}) per iteration.¹² To speed up, we filter bulk of the points beyond a certain distance of a given hyper-rectangle, since its refined hyper-ellipsoid would mostly include/exclude points inside and nearby it. Filtering takes $O(m)$, after which we solve the SDP for a near-constant number of points. It is easy to show that finding the Pareto frontier set of non-dominating packs (line 14)—such that no pack that has strictly larger mass *and* smaller impurity exists—can be done through two passes over all alternative hyper-ellipsoids generated for different (α, λ) . This procedure does not change the overall complexity but is likely to yield a much smaller set of ellipsoids per rectangle. We refine each hyper-rectangle independently in parallel.

The main computation in the last part is the RANDOM-GREEDY algorithm, which makes K iterations for a given number of packs K . In each iteration, it makes a pass over the not-yet-selected hyper-ellipsoids, computes the marginal reduction in bits by selecting each, and picks randomly among the top K with the highest reduction. We use a size- K min-heap to maintain the top K as we make a pass over the packs. Worst case cost is $O(|\mathcal{E}| \log K)$, multiplied by K iterations. We run RANDOM-GREEDY for $K = 1, \dots, a$, each of which is parallelized. Total complexity of §2.3.3 is $O(|\mathcal{E}| a \log a)$.

The number of ellipsoids, $|\mathcal{E}|$, is in the same order of the number of hyper-rectangles from §2.3.1, i.e., $O(c^{d_{\max}})$. Thus, the overall complexity can be written as $O(md_{\max} + c^{d_{\max}} a \log a)$; linear on the number of data points m , near-linear on a , and exponential in the largest pack dimensionality d_{\max} .

2.4 Experiments

Through experiments on real-world datasets we answer the following questions. A quick reference to the UCI datasets used in our experiments is in Table 2.1. Last column gives % savings (in bits) in describing/encoding the anomalies by χ -PACS.

- Q1. **Effectiveness:** How accurate, interpretable, and succinct are our explanations? How do they compare to descriptions by Decision Trees?
- Q2. **Detection performance:** Do our explanations generalize? Can they be used as signatures to detect future anomalies? To this end, we compare χ -PACS to 7 different baselines.
- Q3. **Scalability:** How does χ -PACS's running time scale in terms of data size and dimensionality?

2.4.1 Effectiveness of Explanations

Our primary focus is anomaly *description* where we unearth interpretable characteristics for known anomalies. To this end, we present 6 case studies with ground truth, followed by quantitative comparison to decision trees.

¹²In practice, the solver converges in 20-100 iterations.

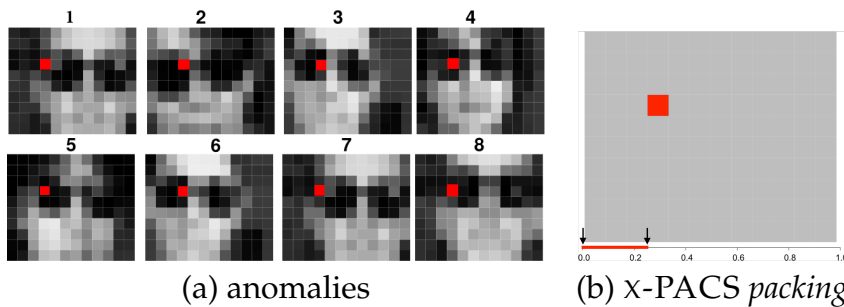
TABLE 2.1: Dataset statistics. χ -PACS achieves significant savings (in bits) by explaining anomalies in groups.

Name	size m	dim. d	anom. a	%-savings
ImagesI	88	120	8	99.75
ImagesII	91	180	9	88.53
ImagesIII	110	180	12	99.51
DigitI	1371	16	228	99.83
DigitII	1266	16	211	99.72
BrCancer	683	9	239	93.74
Arrythmia	332	172	87	92.92
Wine	95	13	24	97.04
Yeast	592	8	129	98.04

Case Studies

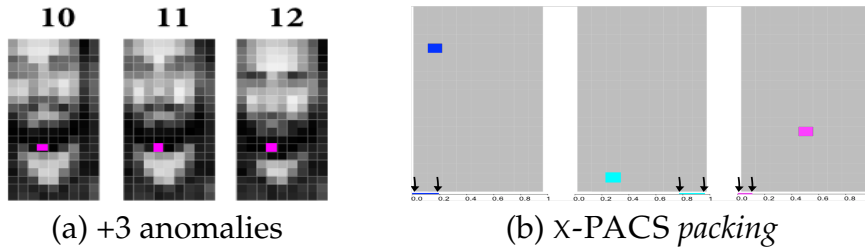
Our **Image dataset** contains gray-scale headshot images of various people. We designate the majority wearing dark-color t-shirts as the normal samples. We create 3 versions containing different number of anomalous patterns, as we describe below. We compare χ -PACS’s findings to the ground truth.

Case I: ImagesI We label 8 images of people wearing sunglasses as anomalies as shown in (a) below, and combine them with the normal samples none of which has sunglasses. In this simple scenario χ -PACS successfully identifies a single, 1-d pattern shown in (b), which packs all the 8 anomalies but no normal samples. Also shown at the bottom of (b) is the interval of values, that is the \pm radius range around the *pack*’s center, for the corresponding dimension (the lower, the darker the pixel).



Case II: ImagesII Next, we construct the 9 anomalies as shown earlier in §2.2 in Fig. 2.1: 6 wearing sunglasses, 4 white t-shirt (2 wearing both), plus 1 person with a beard (normal samples has no beard). As detailed in the caption of the figure, χ -PACS finds 2 pure *packs*, each 1-d, that collectively describe the 8 anomalies and none of the normal samples. The bearded image does not belong to any *pack* and is left out as an outlier.

Case III: ImagesIII We construct the third dataset with 12 anomalies: the same 9 from ImagesII plus 3 faces (10–12) with beard as shown below. In this case, χ -PACS finds that characterizing the bearded images as a separate pattern is best to reduce the description cost, and outputs 3 pure, 1-d *packs* shown in (b).



In all scenarios, x -PACS is able to unearth simple (low-dimensional) and pure (discriminative) characteristics of the anomalies. Also, it automatically identifies the correct number of anomalous patterns that yield the shortest data description as shown in Fig. 2.4.

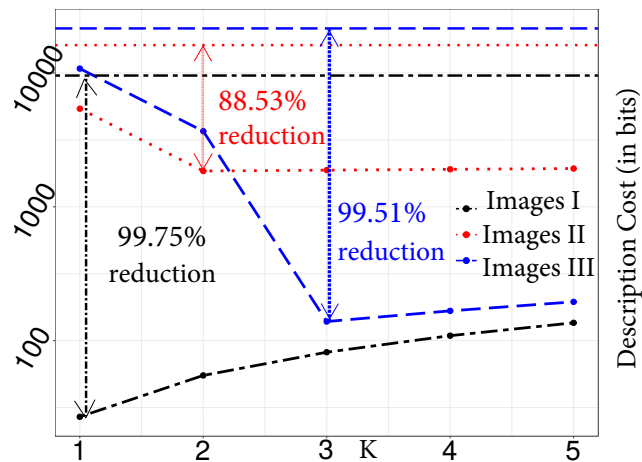
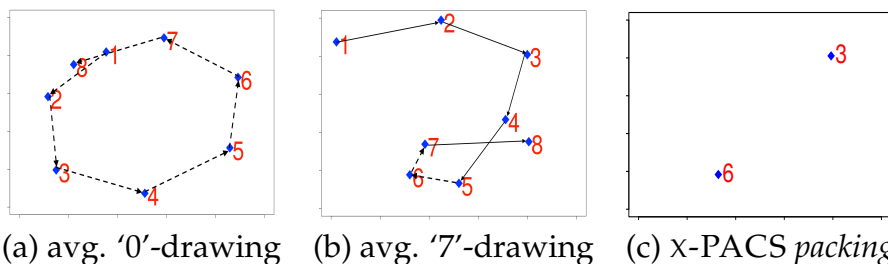


FIGURE 2.4: x -PACS’s description cost of anomalies in image datasets for $K = 1, \dots, 5$. Naïve/base cost ($K = 0$) is shown w/ a horizontal line per dataset. x -PACS finds the appropriate number of patterns automatically and significantly reduces the description cost.

Next we study a different domain. The **Digit dataset** contains instances of digit hand-drawings in time. Features are the x and y coordinates of the hand in 8 consecutive time ticks during which a human draws each digit on paper. As such, each drawing has 16 features.

Case IV: DigitI We designate all drawings of digit ‘0’ as normal and a sample of digit ‘7’ as ‘anomalous’ to study the characteristics of drawing a ‘7’ as compared to a ‘0’. 8 different positions of the hand in time averaged over all corresponding samples of these two digits is shown below (a–b).



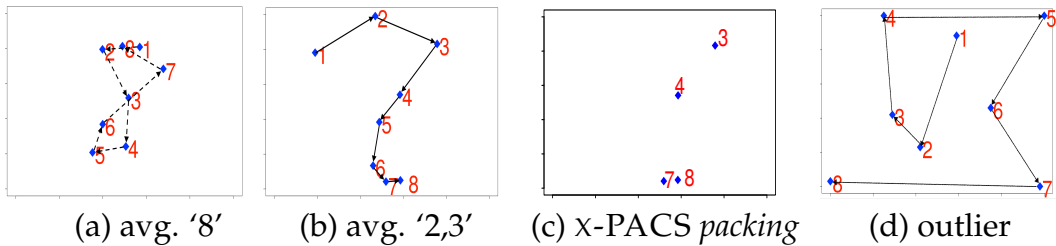
x -PACS identifies a single, 2-d *pack* containing all 228 instances of ‘7’s and no ‘0’s, as given in Table 2.2, where we list the ellipsoid center and the \pm radius interval where the hand is positioned for the characterizing features.

The anomalous pattern suggests right & bottom positioning of the hand respectively at times t_3 & t_6 , which follows human intuition—in contrast, typical hand positions for ‘0’ at those ticks are opposite; at the left & top. Corresponding avg. hand positions in 2-d is shown in (c) above.

TABLE 2.2: DigitI ‘0’ vs. ‘7’: χ -PACS finds one 2-d pack.

packID	feature	center	interval	$ \mathcal{A}_k $	$ \mathcal{N}_k $
$k = 1$	$x@t_3$	0.82	(0.66, 0.98)	228	0
	$y@t_6$	0.17	(0.02, 0.31)		

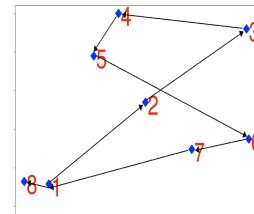
Case V: DigitII We perform a second case study where we designate digit ‘8’ drawings as normal and ‘2’ and ‘3’ as the anomalies. Avg. drawings are illustrated in (a–b) below. χ -PACS is able to describe 210 of the 211 anomalies in a single, 4-d pack listed in Table 2.3 and illustrated in (c). The single unpacked drawing is shown in (d) and looks like an odd ‘3’.

TABLE 2.3: DigitII ‘8’ vs. ‘2’, ‘3’: χ -PACS’s one 4-d pack.

packID	feature	center	interval	$ \mathcal{A}_k $	$ \mathcal{N}_k $
$k = 1$	$y@t_3$	0.83	(0.71, 0.95)	210	0
	$y@t_4$	0.54	(0.38, 0.69)		
	$y@t_7$	0.04	(0.00, 0.11)		
	$y@t_8$	0.05	(0.00, 0.12)		

Looking at the avg. ‘8’ vs. ‘2’ or ‘3’ drawings above, it appears that a single feature like $y@t_8$, i.e., vertical hand position at the end, should be discriminative alone; as ‘8’ tends to end at the *top* vs. others at the *bottom*.

Interestingly, none of the 1-d packs on $y@t_8$ is pure like the 4-d one output. A non-anomalous sample it contains is shown on the right, which is an ‘8’ that starts and ends at the *bottom* just like most ‘2’ and ‘3’s.



Case VI: BrCancer Finally, **breast cancer dataset** contains 239 malign (anomalous) and 444 benign cancer instances. χ -PACS finds 5 packs listed in Table 2.4, covering a total of 226 anomalies while also including 17 unique normal points in the packing. Pack 1 characterizes 162 cases with high ‘chromatin’. Second 2-d pack suggests large ‘clumpthickness’ and ‘mitoses’ (related to cell division and tissue growth) for 145 cases. Smaller pure 1-d packs, 4 and 5, indicate very large ‘cellsize’ and ‘nucleoli’. These findings are intuitive even to non-experts like us (although we lack the domain expertise to interpret pack 3).

¹³Note that RuleFit is averaged over seven datasets due to underspecified regression in Arrhythmia and Yeast

TABLE 2.4: BrCancer: x-PACS finds five 1-d or 2-d packs.

packID	feature	center	interval	$ \mathcal{A}_k $	$ \mathcal{N}_k $
$k = 1$	chromatin	0.76	(0.63, 0.88)	162	11
$k = 2$	clumpthickness	0.94	(0.84, 1.00)	145	5
	mitoses	0.28	(0.00, 0.63)		
$k = 3$	epicellsize	0.33	(0.24,0.42)	97	2
	bareuclei	0.11	(0.09,0.14)		
$k = 4$	nucleoili	0.98	(0.93, 1.00)	75	0
$k = 5$	cellsize	0.99	(0.98, 1.00)	67	0

TABLE 2.5: Interpretability measures (a)–(d): x-PACS vs. Rule learners. Also given for reference is detection performance in AUPRC (See §2.4.2 for details).

measure /method	(a) # of groups	(b) avg. length	(c) avg. impurity	(d) avg. width	detection performance
DT-5	4.0000	2.9889	0.0233	0.4769	0.6252
DT-4	3.7778	2.7856	0.0422	0.4801	0.6070
DT-3	3.0000	2.4078	0.0700	0.4812	0.6210
DT-2	2.4444	1.8889	0.1378	0.4872	0.6236
DT-1	1.7778	1.0000	0.4056	0.5017	0.5656
RuleFit ¹³	12.0000	1.7800	0.0229	0.4643	0.8471
Ripper	2.4444	1.5244	0.0178	0.3889	0.7244
x-PACS	2.5556	2.1000	0.0152	0.2333	0.8781

x-PACS vs. Rule Learners

Since in our work, we view anomalies as an already defined class, explaining anomalies is equivalent to describing an under represented target class (Wrobel, 1997). Hence, we compare x-PACS to techniques that explain labeled data. To this end, we consider interpretable supervised models, specifically, inductive rule based learners that aim to extract rules from a labeled data set that are discriminative in nature. We argue that linear classifiers like logistic regression are not comparable to x-PACS for two key reasons. First, they *do not group* the anomalies, but rather output a single separating hyperplane. Second, they *do not provide rules on the features*, but only feature coefficients, which could be negative (hard to interpret). Further, techniques aiming to explain black box predictions are not directly comparable to our method since most of the works aim to explain one instance at a time compared to the group wise explanations x-PACS provides.

We compare x-PACS to the following popular rule based learners.

1. Decision Tree (DT): DT aims to partition (or group) the labeled data into pure leaves. We treat the leaves containing at least two anomalies

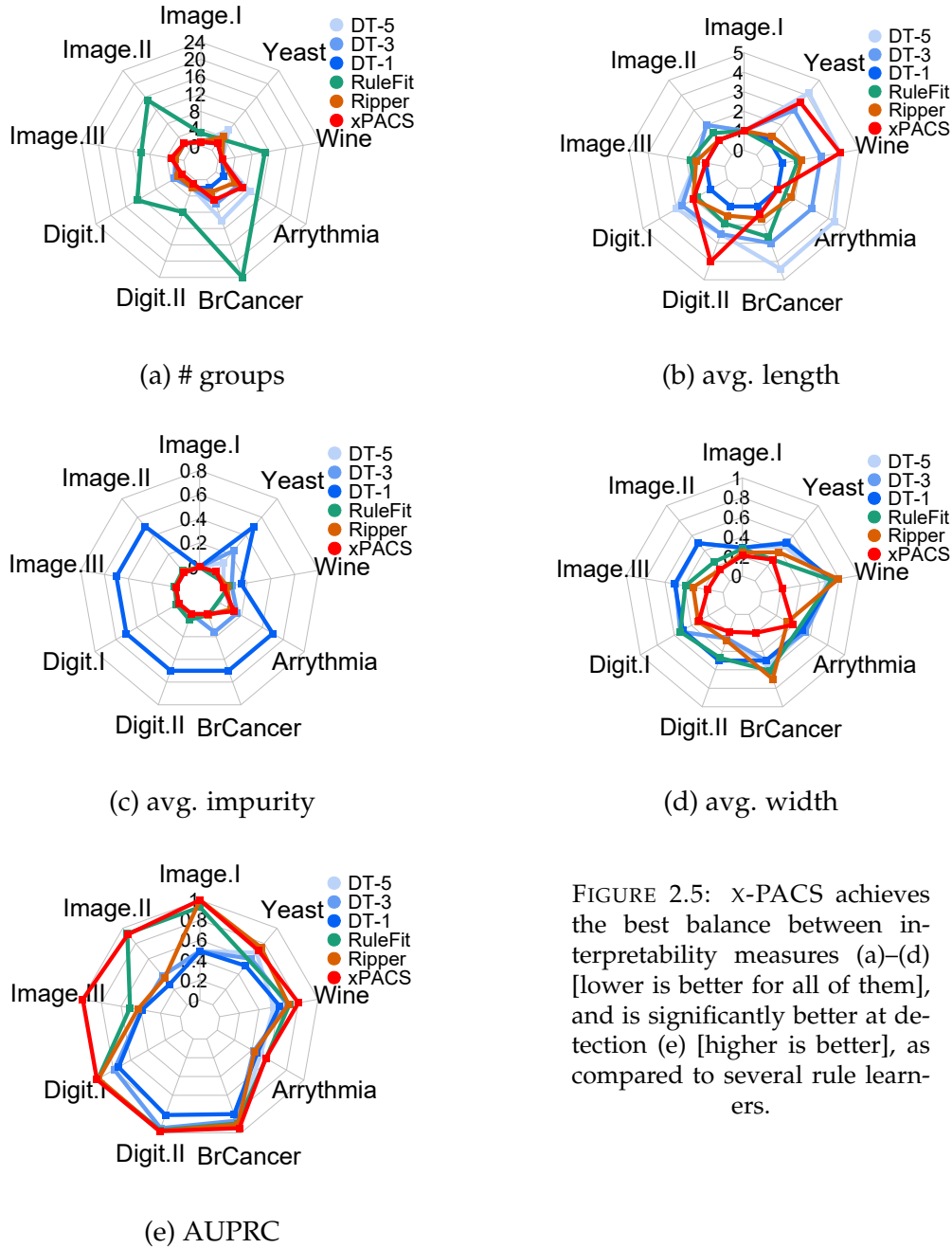


FIGURE 2.5: x-PACS achieves the best balance between interpretability measures (a)–(d) [lower is better for all of them], and is significantly better at detection (e) [higher is better], as compared to several rule learners.

analogous to our *packs*. Each such leaf is characterized by the feature rules (or predicates) on the path from the root.

2. Ripper (Cohen, 1995): Ripper is a popular inductive rule learner that sequentially mines for feature rules with high accuracy and coverage with the aim to achieve generalization. We use a publicly available implementation of Ripper in the Weka repository for our experiments and consider rules that are labeled anomalous.
3. RuleFit (Friedman, Popescu, et al., 2008): RuleFit is an ensemble learner where the base learner is a rule generated by a decision tree. A regression/classification is setup using the base learners to identify the rules that are important in discriminating the different classes. We use the

publicly available RuleFit¹⁴ implementation and use the rules with non-zero coefficients with atleast two anomalies.

To compare χ -PACS with rule learners, it is not fair to use description length since the listed techniques do not explicitly optimize it. Instead, we use the following external interpretability measures proposed in (Lakkaraju et al., 2017) (all being lower the better): (a) number of groups (anomalous packs), (b) avg. length of rules (pack dimensionality), (c) avg. fraction of normal points within packs (impurity divided by n), and (d) avg. interval width across feature rules. In other words, an explanation with fewer groups, fewer rules, fewer exceptions, and smaller spread in features is considered more interpretable.

DT has no means to choose the number of packs automatically. Therefore, we report DT results for depths 1–5 as compared to χ -PACS in Table 2.5, averaged across all datasets. In addition to the interpretability measures, we report the detection performance in AUPRC (area under precision-recall curve) on held-out data (80-20 split) that quantifies the generalization of the subspace rules. Results on individual datasets per measure are shown with radar charts in Fig. 2.5. (See Table 2.6 for detailed results.) Notice the trade-offs between the measures for DT: while (c) and (d) tend to decrease with increasing depth, (a) and (b) increase. The lack of rule summarization in RuleFit is evident in the number of groups (a) where χ -PACS consistently produces smaller number of explanations across various data sets. We also note that χ -PACS produces tighter intervals (d) compared to Ripper emphasizing the concreteness of the explanations. Overall, χ -PACS achieves the best trade-off with lower overall values across the interpretability measures. Moreover, our signatures are significantly better at detecting future anomalies. We present more detailed experiments on detection next.

Ablation Study

We study the importance of the refinement step discussed in §2.3.2 by performing an ablation study. To this end, we omit the refinement of hyper-rectangles into hyper-ellipsoids in χ -PACS (denoted as ablated χ -PACS). Recall that the primary reason we perform the refinement step is to cover more anomalous points and reduce the number of normal points in the packs (See Fig. 2.3). Hence, to showcase the benefit, in Table 2.7, we compare the proportion of anomalous (higher is better) and normal points (lower is better) covered in the final packs obtained using χ -PACS and the ablated χ -PACS. In addition, we also report the %-savings (higher is better) achieved in both cases. In χ -PACS, the summarization step (See §2.3.3) transmits the center and the diagonal matrix of the packs §2.3.3. To accommodate hyper rectangles, we modify this to transmit the upper and lower bounds of the hyper rectangles in the ablated χ -PACS.

From Table 2.7, we observe that χ -PACS is indeed able to cover more anomalous points, while avoiding normal points in the final packs for all the datasets. These results demonstrate the utility of the refinement step.

¹⁴R package pre : <https://CRAN.R-project.org/package=pre>

TABLE 2.6: Rule learners and DT (with respective depths 1–5) compared to χ -PACS across datasets on interpretability measures (a)–(d) [all lower the better] as well as detection performance AUPRC [higher the better]. RuleFit leads to underspecified regression in Arrhythmia and Yeast which we denote by NA.

Measure	Dataset/ Model	Image I	Image II	Image III	Digit I	Digit II	Br Cancer	Arry thmia	Wine	Yeast
(a) number of groups	DT-5	1.00	2.00	2.00	3.00	2.00	10.00	9.00	1.00	6.00
	DT-4	1.00	2.00	2.00	3.00	2.00	8.00	9.00	1.00	6.00
	DT-3	1.00	2.00	2.00	3.00	2.00	6.00	6.00	1.00	4.00
	DT-2	1.00	2.00	2.00	3.00	2.00	4.00	4.00	1.00	3.00
	DT-1	1.00	2.00	2.00	2.00	2.00	2.00	2.00	1.00	2.00
	RuleFit	3.00	15.00	10.00	13.00	8.00	24.00	NA	11.00	NA
	Ripper	1.00	2.00	2.00	2.00	2.00	3.00	5.00	1.00	4.00
	χ -PACS	1.00	2.00	3.00	1.00	1.00	5.00	7.00	1.00	2.00
(b) avg. length of rules	DT-5	1.00	2.00	1.50	3.00	2.50	4.40	4.33	4.00	4.17
	DT-4	1.00	2.00	1.50	3.00	2.50	3.63	3.78	4.00	3.67
	DT-3	1.00	2.00	1.50	2.67	2.50	3.00	3.00	3.00	3.00
	DT-2	1.00	2.00	1.50	2.00	2.50	2.00	2.00	2.00	2.00
	DT-1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	RuleFit	1.00	1.40	1.80	1.80	1.92	2.67	NA	1.81	NA
	Ripper	1.00	1.00	1.50	2.00	1.50	1.67	1.80	2.00	1.25
	χ -PACS	1.00	1.00	1.00	2.00	4.00	1.40	1.00	4.00	3.50
(c) avg. fraction of normal points	DT-5	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.11
	DT-4	0.00	0.00	0.00	0.00	0.00	0.11	0.11	0.00	0.16
	DT-3	0.00	0.00	0.00	0.00	0.00	0.16	0.16	0.07	0.24
	DT-2	0.00	0.00	0.00	0.32	0.01	0.25	0.25	0.08	0.33
	DT-1	0.00	0.50	0.50	0.50	0.50	0.50	0.50	0.15	0.50
	RuleFit	0.00	0.02	0.01	0.02	0.05	0.01	NA	0.05	NA
	Ripper	0.00	0.01	0.00	0.00	0.00	0.01	0.10	0.03	0.01
	χ -PACS	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.00	0.01
(d) avg. interval width	DT-5	0.29	0.50	0.49	0.53	0.26	0.56	0.53	0.68	0.46
	DT-4	0.29	0.50	0.49	0.53	0.26	0.54	0.53	0.68	0.51
	DT-3	0.29	0.50	0.49	0.53	0.26	0.50	0.54	0.72	0.51
	DT-2	0.29	0.50	0.49	0.51	0.29	0.50	0.50	0.79	0.52
	DT-1	0.29	0.50	0.50	0.50	0.50	0.50	0.50	0.73	0.50
	RuleFit	0.28	0.25	0.39	0.53	0.47	0.61	NA	0.72	NA
	Ripper	0.24	0.15	0.31	0.32	0.28	0.7	0.33	0.79	0.38
	χ -PACS	0.21	0.15	0.16	0.31	0.19	0.20	0.39	0.21	0.28
AUPRC	DT-5	0.49	0.38	0.39	0.79	0.95	0.89	0.50	0.55	0.68
	DT-4	0.49	0.38	0.39	0.79	0.95	0.86	0.50	0.55	0.55
	DT-3	0.49	0.38	0.39	0.79	0.95	0.87	0.42	0.69	0.60
	DT-2	0.49	0.38	0.39	0.79	0.95	0.86	0.47	0.61	0.67
	DT-1	0.49	0.27	0.39	0.74	0.81	0.80	0.46	0.61	0.51
	RuleFit	0.93	0.93	0.51	0.98	0.97	0.90	NA	0.71	NA
	Ripper	1.00	0.35	0.43	0.97	0.98	0.91	0.43	0.70	0.75
	χ -PACS	1.00	0.92	0.99	0.99	0.98	0.95	0.56	0.80	0.71

2.4.2 Detection Performance

While not our primary focus, χ -PACS can also be used to *detect* anomalies. Specifically, given the *packs* identified from historical/training data, a future

TABLE 2.7: Ablation Study: x-PACS vs. ablated x-PACS (no refinement to ellipsoids). Coverage of anomalous points (higher is better), coverage of normal points (lower is better), and % savings (higher is better).

	Method	Images I	Images II	Images III	Digit I	Digit II	Br Cancer	Arrythmia	Wine	Yeast
Coverage of anom. points	x-PACS	1.00	0.89	1.00	1.00	1.00	0.95	0.93	0.96	0.68
	ablated x-PACS	1.00	0.89	1.00	0.96	0.89	0.88	0.78	0.92	0.62
Coverage of normal points	x-PACS	0	0	0	0	0	0.03	0.35	0.11	0.10
	ablated x-PACS	0	0	0	0.01	0.01	0.05	0.53	0.18	0.13
%savings	x-PACS	99.75	88.53	99.51	99.83	99.72	93.74	92.92	97.04	98.04
	ablated x-PACS	99.75	88.53	99.51	92.11	87.21	85.68	78.16	91.42	90.51

test instance that falls in any one of the *packs* (i.e., enclosed within any hyper-ellipsoid in the *packing*) can be flagged as an anomaly.¹⁵

To measure detection quality, we compare x-PACS to 7 competitive baselines on all datasets.

1. Mixture of K -GAUSSIANS on the anomalous points. $K \in \{1, 2, \dots, 9\}$ chosen at the “knee” of likelihood. Anomaly score of test instance: maximum of the probabilities of being generated from each cluster.
2. KDE on the normal points. Gaussian kernel bandwidth chosen by cross-validation. Anomaly score: negative of the density at test point.
3. NN. Anomaly score: distance of test point to its nearest neighbor (nn) normal point in training set, divided by the distance of that nn point to its own nearest normal point in training set.
4. PCA+SVDD on all points (Tax and Duin, 2005). A *single* hyperball that aims to enclose anomalous points in the *PCA-reduced* space¹⁶, for which the embedding dimensionality is chosen at the “knee” of the scree plot. Anomaly score: distance of test point from the hyperball’s center.
5. DT on all points, where we balance the data for training and regularize by tree-depth, chosen from $\{1, 2, \dots, 30\}$ via cross-validation. Anomaly score: number of anomalous samples in the leaf the test point falls into divided by leaf size.
- 6-7) SVM-LIN & SVM-RBF on all points. Hyperparameters set by cross-validation. Anomaly score: “confidence”, i.e., distance from decision boundary.

We create 3 folds of each dataset, and in turn use 2/3 for training and 1/3 for testing, except the Images datasets with the fewest anomalies for which we do leave-one-out testing. All points receive an anomaly score by each

¹⁵Note that, like any supervised method, x-PACS could only detect future instances of anomalies of known types.

¹⁶SVDD optimization diverged for some high dimensional datasets, therefore, we performed PCA as a preprocessing step.

TABLE 2.8: Area under precision-recall curve (AUPRC) on anomaly detection.

Method	ImagesI	ImagesII	ImagesIII	DigitI	DigitII	BrCancer	Arrythmia	Wine	Yeast
K-GAUSSIANS	0.182	0.239	0.184	0.162	0.333	0.613	0.227	0.258	0.265
KDE	0.952	0.978	0.987	0.989	0.997	0.981	0.571	0.667	0.681
NN	0.491	0.472	0.659	0.967	0.821	0.520	0.546	0.562	0.348
PCA+SVDD	0.286	0.217	0.212	0.331	0.529	0.861	0.295	0.566	0.606
DT	0.802	0.764	0.812	0.831	0.961	0.884	0.516	0.637	0.673
SVM-LIN	1.000	1.000	1.000	0.999	0.999	0.984	0.755	0.994	0.823
SVM-RBF	1.000	1.000	1.000	1.000	1.000	0.964	0.810	0.984	0.861
χ -PACS	1.000	0.921	0.990	0.993	0.976	0.951	0.564	0.799	0.701

method as described above. χ -PACS’s anomaly score for a test instance \mathbf{x} is the maximum $h_k(\mathbf{x}) = \mathbf{x}^T \mathbf{U}_k \mathbf{x} + \mathbf{w}_k^T \mathbf{x} + w_{0k}$ among all p_k ’s in the *packing* resulting from training data. We rank points in decreasing order of their score, and report the area under the precision-recall curve in Table 2.8.

SVMs achieve the highest detection rate, as might be expected. However, kernel SVM cannot be interpreted. Linear SVM, like LR, does not identify anomalous patterns nor does it produce any explicit feature rules. Notably, χ -PACS outperforms all other baselines considerably across datasets, including DT, which produces the most interpretable output among the baselines as discussed in §2.4.1.

2.4.3 Scalability

Finally, we quantify the scalability of χ -PACS empirically. To this end, we implement a synthetic data generator, parameterized by data size, total dimensionality, maximum pack size and dimensionality and number of anomalous packs. Anomalies are sampled from a small range per feature within a subspace, and normal points are sampled from the reverse of the histogram densities derived from the anomalous points.

Fig. 2.6 shows the running time w.r.t. data size m , dimensionality d , average pack dimensionality d_{avg} , and total number of anomalies a . All plots demonstrate near-linear scalability. Recall that we showed exponential complexity w.r.t. maximum pack dimensionality d_{max} . Notably, we observe linear time growth on average.

2.5 Related Work

Related areas of study span across outlier explanation, subspace clustering and subspace outlier detection, data description, subgroup discovery, rule learning, rare class discovery and methods to explain black-box classifiers. We illustrate related work in the context of our desiderata in Table 2.9.

Outlier explanation: The seminal work by Knorr and Ng (Knorr and Ng, 1999) provides what they call “intensional knowledge”, per outlier, by identifying the minimal subspaces in which it deviates. To find the optimal subset of features that differentiate the outliers from normal points, (Kuo and

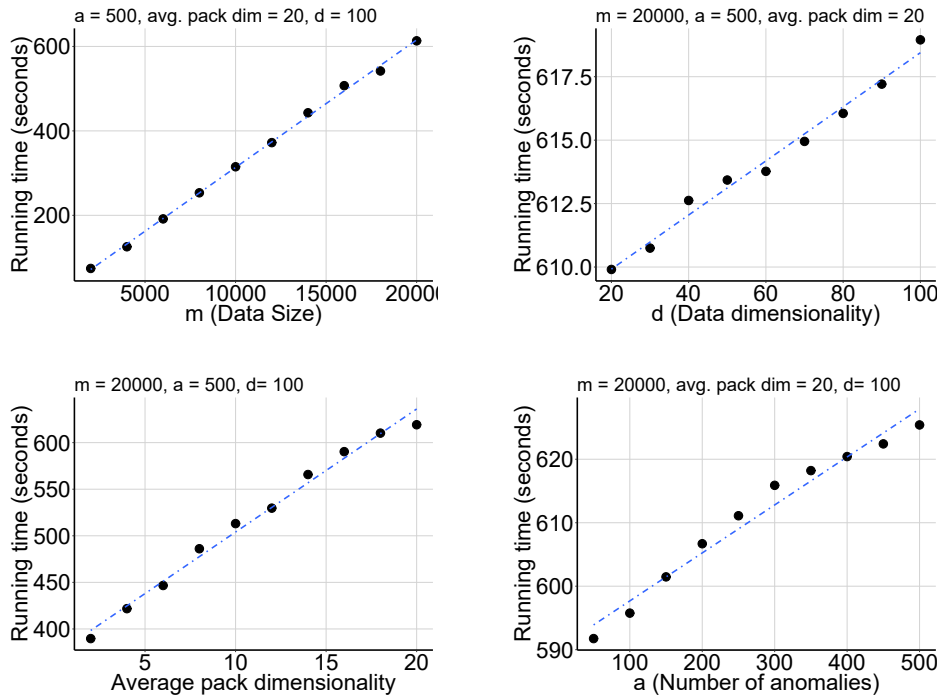


FIGURE 2.6: x-PACS scales linearly with input size.

Davidson, 2016) formulates a constraint programming problem and (Keller et al., 2013) takes a subspace search route. Similarly, (Dang et al., 2014; Dang et al., 2013; Micenková et al., 2013) aim to explain one outlier at a time by features that participate in projection directions that maximally separate them from normal points. All existing work in this area assume the outliers are *scattered* and strive to explain them *individually* rather than in groups. Therefore, they cannot identify anomalous patterns. Moreover, they do not focus explicitly on shortest description, let alone in a principled, information-theoretic way as we address in this work.

Extending earlier work (Angiulli, Fassetti, and Palopoli, 2009) on explaining single outliers, (Angiulli, Fassetti, and Palopoli, 2013) aims to explain groups of outlier points or what they call sub-populations. They search for $\langle \text{context}, \text{feature} \rangle$ pairs, where the (single) feature can differentiate as many outliers as possible from normal points that share the same context. It is important to note that their goal is to explain a group (or set) of outliers altogether and not particularly explaining them with multiple groups. Similarly, (Zhang, Diao, and Meliou, 2017) describes anomalies grouped in time. They construct explanatory Conjunctive Normal Form rules using features with low segmentation entropy, which quantifies how intermixed normal and anomalous points are. They heuristically discard highly correlated features from the rules to get minimal explanations. Again, they strive to explain all the anomalies as a group, and not in multiple groups.

We found that SRF (sapling random forest) (Kopp, Pevný, and Holena, 2014) aims to explain and cluster outliers similar to our problem setting. They build on their earlier work (Pevný and Kopp, 2014), which explains outliers one at a time by learning an ensemble of small decision trees (called saplings) and combining the rules (from root to leaf in which the outlier

lies) across the trees. SRF then groups the outliers using k-means clustering based on the similarity of their explanations. However, there is no guarantee on the minimality of their overall description, since grouping is done as a post-processing step and by using a local-optima-prone clustering algorithm. Moreover, there is not much discussion in their paper on the choice of the number of clusters, nor the format of the final description after the anomalies are clustered. We are not aware of a publicly available implementation of SRF to compare with our proposed method and hence omit it from the experimental evaluation.

Subspace clustering and Outlier detection: There is a long list of work on subspace clustering (Agrawal et al., 1998; Cheng, Fu, and Zhang, 1999; Kriegel et al., 2005; Müller et al., 2009; Sequeira and Zaki, 2004) that aim to find high-density clusters in feature subspaces. (See (Parsons, Haque, and Liu, 2004; Kriegel, Kröger, and Zimek, 2009) for reviews.) Some others are projection based that work in transformed feature spaces (Aggarwal et al., 1999; Moise, Sander, and Ester, 2006). However, these are unsupervised methods and their goal is not explaining labeled data, nor they focus on minimal explanations. There is also a long list of subspace-based outlier *detection* methods (Keller, Müller, and Böhm, 2012; Kriegel et al., 2009; Kriegel et al., 2012; Müller et al., 2012; Müller et al., 2008; Müller, Schiffer, and Seidl, 2011), however, those do not address the description problem.

Data description and Rare class discovery: Another line of related work is data description (Görnitz, Kloft, and Brefeld, 2009; Tax and Duin, 2005) and rare class (or category) characterization (He, Tong, and Carbonell, 2010; He and Carbonell, 2010). The main goal behind all of these work is to explain the data via a hyperball separating rare-class points from normal points. However, all of them assume that the former cluster in a *single* hyperball, and with the exception of (He and Carbonell, 2010), search for a *full-dimensional* enclosing hyperball. As such, they do not address the curse-of-dimensionality or identify multiple clusters embedded in different subspaces.

Subgroup discovery and Rule learning: Classification rule learning algorithms have the objective of generating models consisting of a set of rules inducing properties of all the classes of the target variable, while in subgroup discovery the objective is to discover individual rules of interest (See (Herrera et al., 2011) for an overview). The seminal work of rule based learners such as Ripper (Cohen, 1995) and CN2 (Clark and Niblett, 1989) sequentially mine for rules with high accuracy and coverage. More recently, (Friedman, Popescu, et al., 2008) proposes RuleFit, an ensemble learner where the base learner is a rule generated by a decision tree. A regression/classification is setup using the base learners to identify the rules that are important in discriminating the different classes. Few other work in ensemble learning (Deng, 2014; Hara and Hayashi, 2016) build ensemble trees. While rules are interpretable, they are learnt with an aim to achieve generalization (to unseen data). This is different from our work where we primarily focus on describing the under-represented class (anomalies) without emphasizing the generalization.

SubgroupMiner (Klösgen and May, 2002) extends seminal work in subgroup discovery (MIDOS (Wrobel, 1997), Explora (Klösgen, 1996)) to handle

numerical and categorical attributes. SD (Gamberger and Lavrac, 2002) proposes an interactive subgroup discovery technique based on the variation of beam search algorithms guided by expert knowledge. Krimp (Vreeken, Van Leeuwen, and Siebes, 2011) proposes a greedy MDL based approach to mine few frequent item sets describing a dataset. This method can be used on multi-class data to provide descriptions of each class, it is however limited to categorical attributes. Discriminative pattern mining techniques (Loekito and Bailey, 2008) also assume categorical features and aim to extract contrast patterns (item sets) with large support difference across classes.

A key difference between the techniques discussed above and our work is the summarization scheme described in §2.3.3. Our MDL based encoding scheme leads to a submodular rule selection with theoretical guarantees that the current subgroup discovery or rule learning algorithms do not explore. The improvement of the summarization scheme is evident from our experiments (See Table 2.6), comparing χ -PACS to several rule learners on various interpretability measures.

Explaining black-box classifiers: Approaches such as (Fong and Vedaldi, 2017; Koh and Liang, 2017; Montavon, Samek, and Müller, 2017; Ribeiro, Singh, and Guestrin, 2016) aim to explain the decision made by a black-box predictor. LIME (Ribeiro, Singh, and Guestrin, 2016) finds nearest neighbors to single input labeled example to construct a linear interpretable model that is locally faithful to the predictor. Further, authors propose a submodular optimization framework to pick instances that are representative of the predictions of a classifier. Other work (Fong and Vedaldi, 2017; Koh and Liang, 2017) explain the model by perturbing the features to quantify the influence on prediction. However, these work do not aim to explain multiple instances collectively, as such they do not handle the summarization problem, which are hence not comparable to our proposed method.

All in all, none of the existing methods provides all of 1) *collective* (as opposed to individual) explanations, 2) explanations for *multiple* anomalous groups, 3) in characterizing *subspaces*, 4) using interpretable *feature rules* that can 5) *discriminate* anomalies from normal points, 6) aiming to *minimize* description length.

2.6 Conclusion

We considered the problem of explaining given anomalies in high-dimensional datasets in groups. Our key idea is to describe the data by the patterns it contains. We proposed χ -PACS for identifying a small number of low-dimensional anomalous patterns that “pack” similar, clustered anomalies and “compress” the data most succinctly. In designing χ -PACS, we combined ideas from data mining (bottom-up algorithms with pruning), optimization (nonlinear quadratic discrimination), information theory (data encoding with bits), and theory of algorithms (nonmonotone submodular function maximization). Our notable contributions are listed as follows.

- **A new desiderata** for the anomaly description problem, enlisting five desired properties (D1–D5),

- **A new problem formulation**, for explaining a given set of anomalies *in groups* (D1),
- **Description algorithm** χ -PACS, which provides low-dimensional (D2), interpretable (D3), and discriminative (D4) feature rules per anomalous group,
- **A new anomaly encoding scheme**, based on the minimum description length (MDL) principle, that lends itself to efficient optimization to produce **minimal explanations** (D5) with guarantees.

Through experiments on real-world datasets, we showed the effectiveness of χ -PACS both in explanation and detection and superiority to competitive baselines. For reproducibility, all of our source code and datasets are publicly released at <https://github.com/meghanathmacha/xPACS>.

Acknowledgments

This research is sponsored by NSF CAREER 1452425 and IIS 1408287, ARO Young Investigator Program under Contract No. W911NF-14-1-0029, and the PwC Risk and Regulatory Services Innovation Center at Carnegie Mellon University. Any conclusions expressed in this material are of the authors and do not necessarily reflect the views, either expressed or implied, of the funding parties.

TABLE 2.9: Comparison of related work in terms of properties D1–D5 in reference to our Desiderata (see §2.1.1).

Property	Explain as groups?	Multiple groups?	Find subspaces?	Rules on features?	Discriminative?	Minimal?
Subspace clustering (Agrawal et al., 1998; Cheng, Fu, and Zhang, 1999; Sequeira and Zaki, 2004; Kriegel et al., 2005; Müller et al., 2009)	✓	✓	✓	✓	✓	✓
Projected clustering (Aggarwal et al., 1999; Moise, Sander, and Ester, 2006)	✓	✓	✓	✓	✓	✓
(data descr.) SVDD (fax and Duin, 2005), SSSVDD (Görnitz, Kloft, and Brefeld, 2009)	✓	✓	✓	✓	✓	✓
(rare category) RACH (He, Tong, and Carbonell, 2010)	✓	✓	✓	✓	✓	✓
(rare category) PALM (He and Carbonell, 2010)	✓	✓	✓	✓	✓	✓
Knorr and Ng (Knorr and Ng, 1999)	✓	✓	✓	✓	✓	✓
RefOUT (Keller et al., 2013), CP (Kuo and Davidson, 2016) LODI (Dang et al., 2013), LOGP (Dang et al., 2014)	✓	✓	✓	✓	✓	✓
EXPRES (Angiulli, Fassetti, and Palopoli, 2013)	✓	✓	✓	✓	✓	✓
(Explaining black box classifiers) LIME (Ribeiro, Singh, and Guestrin, 2016)	✓	✓	✓	✓	✓	✓
Exstream (Zhang, Diao, and Meliou, 2017)	✓	✓	✓	✓	✓	✓
Explainer (Pevný and Kopp, 2014)	✓	✓	✓	✓	✓	✓
SRF (Kopp, Pevný, and Holena, 2014), Krimp (Vreeken, Van Leeuwen, and Siebes, 2011), RuleFit (Friedman, Popescu, et al., 2008), Ripper (Cohen, 1995)	✓	✓	✓	✓	✓	✓
x-PACS [this paper]	✓	✓	✓	✓	✓	✓

Chapter 3

Social Determinants of Health

3.1 Introduction

3.1.1 Social Determinants of Health

The U.S. sees 35.7 million hospital stays per year, representing a hospitalization rate of 104.2 stays per 1,000 population. Hospitalizations incur enormous costs, \$417.4 billion per year and \$11,700 per stay. Since the hospitalization rate increases with age, from 17.1 (1-17 years old) to 455.7 (85+) per 1,000 population (Freeman, Weiss, and Heslin, 2018; Rosenberg et al., 2016), the hospitalization costs will continue to skyrocket with an aging population. Moreover, only 20% of an individual's health is attributable to access to healthcare, whereas 80% to the remaining components of *social determinants of health*: physical environment, socio-economic factors, and lifestyle choices (American Hospital Association (ICSI, 2004)). It is thus imperative to leverage new data, beyond the conventional patient data available to healthcare professionals, and develop new tools to understand factors that predict individual health risk, reduce hospitalization rate, and promote population health.

A variety of literature empirically investigated the importance of the factors beyond the clinical wall on health outcomes. Diet, smoking cessation, exercise, and sleep are shown to critically improve life expectancy and reduce hospitalization costs (cf. (Chen, Tan, and Padman, 2020) for a recent review). Up to one-third of premature deaths in the U.S. arise from conditions modifiable via lifestyle choices (Loewenstein, Brennan, and Volpp, 2007). Socio-economic factors, such as income and education, are associated with life expectancy with the greatest disparities occurring in the mid-adulthood (Greer et al., 2014). These studies primarily rely on identification of social determinants of health from electronic health records (EHR), medical claims, and individual surveys. Deviating from the status quo, this research examines the associations between health outcomes and individual's social determinants of health by leveraging atomic, longitudinal individual smartphone location data.

3.1.2 Location Data

Identifying social determinants of health from mobile location data presents several significant advantages over conventional data sources. First, mobile location data are straightforward to collect, merely an app permission away, tracked in the background in most mobile ecosystems, and readily accessible

to data users. Sustained data collection also requires little to no effort from an individual or data user, compared to a hospital visit or medical claim filing. Second, mobile location data offer an extensive, spatio-temporal profile of an individual by delineating day-to-day behavior, mobility, lifestyle choices, and social relations (Ghose, Li, and Liu, 2018). Meanwhile, these data embed rich points-of-interests (POIs), such as restaurants, gyms, pharmacies, and hospitals, home and work locations (Macha et al., 2019). Third, mobile location data portray a much richer context than EHR, such as the longer-term precursors (i.e., locations visited and behaviors before) and aftermath of a hospital visit. Fourth, mobile location information can help fill any data void (e.g., when no EHR or health insurance is available for a first-time patient) or verify survey responses. Fifth, mobile location data permit continuous monitoring of social determinants of health, thus facilitating adaptive interventions to mitigate future health risk (Wachs et al., 2015). In a nutshell, we aim to propose a framework to identify the social determinants of health from these behaviorally rich individual location data and empirically quantify their association with future health outcomes of immense economic and societal values.

3.1.3 Research Gap

Studies across disciplines have aimed to understand individual behavior from location data – characterizing mobility patterns (Gonzalez, Hidalgo, and Barabasi, 2008), social ties (Morse, Gonzalez, and Markuzon, 2016; Eagle, Pentland, and Lazer, 2009), and shopping patterns (Hu et al., 2016). While most behavioral patterns have been leveraged for advertising (Molitor et al., 2019), their relationship with health outcomes is receiving increased attention. On one hand, researchers primarily from Computer Science focus on identifying macro representations of an individual's subset of daily activities without linking to long-term health risk (Logan et al., 2007; Farrahi and Gatica-Perez, 2011; Sun et al., 2014). On the other, the medical community examines health outcomes, such as depression, schizophrenia symptoms (Ben-Zeev et al., 2014) and other standard clinical measurements (Robben, Pol, and Kröse, 2014), by analyzing micro activities, such as sleep patterns, gait, and activity rhythms. Both literature rely on sensor data from fewer than 200 individuals.

In comparison, our research is distinctive on multiple fronts. We extract a *comprehensive* range of behavioral patterns, including work, leisure, commute, and fitness, to capture "lifestyle", defined in sociology and marketing as "an activity that exhibits a pattern of behavior, consumption or leisure" (Cockerham, Abel, and Lüschen, 1993). We further integrate these macro representations of lifestyle with micro-level features inferred from the location data, such as accessibility to healthcare facilities and socioeconomic status, to construct an extensive profile of an individual's social determinants of health. We then quantify the link between these determinants and a key health outcome - hospitalization. Our examination of the population-scale data also permits empirical generalization and policy guidance.

3.1.4 Overview of Proposed Methodology

We summarize the proposed framework that comprises two key components: Identification of social determinants and Health risk quantification.

Social Determinants : To identify individual lifestyles, we build on unsupervised topic models. Topic models are generative models that represent documents as mixtures of topics, learned in a latent space, and allow for clustering and ranking of documents, words, and other entities, like authors. We identify 16 activity groups (Table 3.1) grounded in sociologist's definition of routine and leverage Author Topic Models (ATM). In ATM, we map the concept of word to an activity combined with a temporal context (e.g., a *restaurant* visit during 9 - 11 AM as *restaurant.911*); document to a bag of activities in a day, individual to an author and successfully identify author-specific macro activity patterns across multiple days – routines. Further, from location data, we identify *home*, *work* locations of an individual, to infer their neighborhood economic stability, social community context, accessibility to resources, and socio-economic factors. (Table 3.2)

Health Risk Quantification : To study if the identified characteristics signal health events, we designate the individual's health event based on their hospital visits in the near future. We perform a model-free analysis to infer population-level association of individual routines with future health events. To quantify the importance of identified individual characteristics in relation to future health, we specify a logit model. Finally, to quantify the health risk, we build on the concepts of multi-modal and sequential deep learning to unify multiple types (refer to Table 3.2 for the full set) of individual characteristics – time varying categorical (e.g. routines), numerical (e.g. frequency of restaurant visits) and static categorical (e.g. workplace of the individual), numerical (e.g. average household income of the census block where the individual lives in) and predict the health event for each individual. We validate our proposed method on locations of over 10,000 individuals from Baltimore and D.C. over four months in 2019.

3.1.5 Key Findings

For Baltimore residents, the lifestyle identification reveals that while as expected the weekday (weekend) lifestyle is primarily characterized by work (home) routines, heterogeneous lifestyles, such as workaholics and fitness regulars, do emerge. For instance, we find that individuals with a late working routine are more likely to consume at restaurants during the night (9 - 11 PM) in contrast to early working routine individuals who prefer going to restaurants during afternoon (2 - 5 PM). Individuals with constant work, limited fitness, or stay at home on weekdays are 2.01 and 1.47 times more likely to have a future hospitalization within the next year compared to average (2.45%). In contrast, those who conduct fitness on weekends or weekdays are much less likely (0.52 and 0.65 times, respectively) to have a hospitalization. Interestingly and importantly, *regularity*, rather than total time spent at healthy activities and unhealthy activities, significantly predicts future hospitalization. Overall, an individual's lifestyle choice is more critical than the socio-economic and accessibility factors.

Finally, to quantify the health risk, we jointly represent the multiple facets of an individual's social determinants and develop a sequential deep learner to predict future hospitalization. The proposed learner, dealing with a huge class imbalance (2.45 % on average are hospitalized) achieves a PR AUC and ROC AUC of 0.28 and 0.85 respectively. From an ablation study of the proposed learner and several baselines, we confirm that individual behavioral features, such as lifestyles and day-to-day activities, significantly contribute to the predictive performance for both Baltimore (16.6% increase in PR AUC) and D.C. residents (30% increase). These findings remain consistent across the proposed learner and considered baselines.

The rest of the manuscript is organized as follows. In Section 3.2, we review literature from various disciplines that are relevant to our research questions. Section 3.3 describes the details of the proposed framework. In Section 3.4, we provide details of our sampling and summary statistics of the mobile location data under analysis. In Section 3.5, we discuss the empirical results and advantages of the proposed framework. We offer concluding remarks in Section 3.6.

3.2 Related Work

We will concisely review the most relevant Marketing, Management, Information Systems (IS), Computer Science (CS) and Medical literature on individual routines, lower level activity recognition and their impact, associations with health events.

3.2.1 Behavioral Routine and Activities:

We break down this stream based on the type of data used in the study.

Smartphone Data

Researchers, primarily from the CS community developed several machine learning techniques to recognize low-level individual activities (e.g., sitting, standing, or walking) and high-level activities, often referred to as *lifestyles* or *routines*, (e.g., eating at a restaurant, taking a subway) from various types of sensor data collected from smartphones. While some of these methods are supervised (Bao and Intille, 2004; Logan et al., 2007), due to the practical limitation of acquiring labeled data for activities, a majority of the recent focus has shifted towards unsupervised methods (Eagle and Pentland, 2009; Farrahi and Gatica-Perez, 2011; Huynh, Fritz, and Schiele, 2008; Sun et al., 2014; Zheng and Ni, 2012). (Eagle and Pentland, 2009) use principal component analysis (PCA), a dimensionality reduction technique, to obtain main components that construct human daily routines. However, the resulting eigenvectors cannot be mapped to a specific activity of an individual (lesser interpretability) and do not capture the temporal nature of individual activities. (Huynh, Fritz, and Schiele, 2008) discover daily routines from wearable sensor data using *K*-means clustering to build activity vocabulary involving activities such as *dinner*, *commuting*, *office* and use LDA to identify routines. (Farrahi and Gatica-Perez, 2011) apply LDA and ATM on labeled

cell tower data to automatically discover routines, including “being at work” or “going home from work”. (Zheng and Ni, 2012) propose a probabilistic generative model for learning individuals’ latent behavior patterns based on unlabeled cell tower data. (Sun et al., 2014) develop a non-parametric framework for human routine discovery using a combination of the Dirichlet Process Mixture Model (DPMM) and Hierarchical Dirichlet Process (HDP). This sub-stream of literature limit their focus to a subset of an individual’s daily activities (such as work or shopping patterns) and do not analyze potential long-term health signals from the identified representations.

Surveys and Health Records

Several other measures of routines have been developed in the medical literature via surveys or individual health records (Guenther, Reedy, and Krebs-Smith, 2008; Chiuve et al., 2012; Joumard et al., 2010). These measures are based on smoking cessations, physical activity, diet quality, alcohol consumption and body weight. Healthy Eating Index-2015 (HEI-2015), computed based on individual surveys is a measure for assessing whether a set of foods aligns with the Dietary Guidelines for Americans (DGA). Alternatives to HEI with stronger correlations to chronic diseases was proposed by (Chiuve et al., 2012). Acquiring longitudinal measures of such nature - for instance, via surveys, would require frequent interaction with individuals making them less practical than smartphone data based inference. Next, we discuss works that study associations and impact of behavioral routines, activities on future health events.

3.2.2 Behavior as Health Determinants:

Individual behavior, measured as dietary, alcohol and tobacco consumption have been studied to determine health status of a population (Joumard et al., 2010). Impact of of lifestyle factors, determined by physical activity, high dietary score AHEI-2010 (Chiuve et al., 2012) on premature mortality was studied by (Li et al., 2018). The study revealed that the projected life expectancy at age 50 years was on average 14 years longer among female Americans with 5 low-risk routine factors (moderate to vigorous physical activity, moderate alcohol intake, and a high diet quality score) compared with those with zero low-risk factors; for men, the difference was 12.2 years. Other factors such as health care resources (Miller and Frech, 2002), socio-economic factors (Nixon and Ulmann, 2006) have been studied to impact health outcomes of a population. This stream of study primarily rely on data from surveys, electronic health records, medical claims differing from our work involving location data.

Prior studies have associated sensor measurements of sleep patterns, gait, activity rhythms, indoor activities and outings, and mobility with standard clinical measurements and survey data. (Paavilainen et al., 2005) compares changes in the circadian rhythm of day to day activities of older adults living in nursing homes with clinical health measurements. Mobility metrics derived from location data have been used to describe the patterns of behavior and subjective experience associated with depressive symptoms (Saeb et

al., 2015), and mood patterns associated with schizophrenia symptoms (Ben-Zeev et al., 2014). (Robben et al., 2012; Robben, Pol, and Kröse, 2014) study relationship between location and transition patterns of an individual's indoor mobility behavior, namely the frequency, duration and times being carried out, with the Assessment of Motor and Process Skills (AMPS) scores (Fisher and Jones, 2012). Other works have explored the relationship between walking speed and the amount of in-home activity among healthy older adults and older adults with Mild Cognitive Impairment (MCI) (Hayes et al., 2008). This study revealed that the coefficient of variation in median walking speed was higher in for adults with MCI group when compared to healthy individuals. Wearable sensor data was used to infer physical activity in patients with knee osteoarthritis (Agarwal et al., 2018). (Dawadi, Cook, and Schmitter-Edgecombe, 2016) introduces the notion of an activity curve, which represents a visual abstraction of an individual's routines and develops a technique to detect changes in routines and perform health assessment. Our work complements this line of literature by identifying and associating routines with future health outcomes from individual location data. In addition, we also present a sequential deep learner to quantify individual health risk. The quantification is based on multiple facets of an individual day-day behaviour comprising of routines, mobility and socio-economic factors. To the best of our knowledge, we are not aware of other works that involve prediction of future health events from location data.

3.3 Framework

The primary objectives of our framework are two-fold. First is to identify an individual's social determinants of health from the location data: such as lifestyles, socioeconomic status, and accessibility to various resources. Second is to quantify the relationship between these determinants and individual health risk. We will introduce the relevant notations next.

Definition 7 (Trajectory) A trajectory T_i of an individual i is defined as a temporally ordered set of tuples $T_i = \{(l_1^i, t_1^i), \dots, (l_{n_i}^i, t_{n_i}^i)\}$, where $l_j^i = (x_j^i, y_j^i)$ is a location where x_j^i and y_j^i are the coordinates of the geographic location¹, and t_j^i is the corresponding time stamp.

Definition 8 (Activity-Trajectory) An activity trajectory D_i of an individual i is defined as mapping T_i to activities that exhibit a pattern of behavior, consumption, leisure. D_i is a temporally ordered set of tuples $D_i = \{d_1^i, \dots, d_{n_i}^i\}$, $d_j^i = (a_j^i, c_j^i)$, where $a_j^i = \text{act}(l_j^i)$, $l_j^i \in T_i$ is an activity by the individual inferred from a location closest to x_j^i and y_j^i , and c_j^i is a coarser timestamp² of t_j^i . Also, denote W as the universe of all temporal activities d_j^i across D_i .

Definition 9 (Routine) A routine L_i of an individual i is defined as a set of activities and their corresponding timestamps $L_i = \{d_1^i, d_2^i, \dots, d_Y^i\}$, $d_j^i = (a_j^i, c_j^i) \in W$,

¹Coordinates usually correspond to latitude and longitude.

²For instance, $t_j^i = 9:33$ AM is coarsened and represented as 9 - 11 AM.

$|L_i| = Y$ that globally represent an individual's day to day temporal activities across T_i .

Next, we illustrate the transformation of individual trajectories (T_i) to activity trajectories (D_i) (Section 3.3.1) and detail the identification of lifestyles (L_i , Section 3.3.2). In Section 3.3.3, we discuss the remaining social determinants and our learner to quantify health risk. We present model-free analysis to understand if lifestyles signal future hospitalizations and discuss the performance of the proposed learner in predicting them in Section 3.5.

Activity group	Place type of location
hospital	hospital, doctor
health	physiotherapist, pharmacy, dentist, drugstore
necessityshopping	store, supermarket, convenience_store, home_goods_store, grocery_or_supermarket, hardware_store
fitness	gym
publictransport	transit_station, train_station, bus_station, light_rail_station, subway_station
owntransport	car_wash, car_repair, parking, gas_station, taxi_stand
religious	church, mosque, hindu_temple, synagogue
recreation	amusement_park, tourist_attraction, zoo, park, theatre, sports_stadium, concert, bowling_alley, art_gallery, aquarium, museum, movie_rental, book_store, library, movie_theater, campground
travel	hotel, lodging, rv
personalcare	beauty_salon, spa, hair_care
leisureshopping	clothing_store, department_store, shopping_mall, shoe_store, electronics_store, furniture_store
unhealthyactivities	casino, liquor_store, bar, night_club, cigarette
restaurant	restaurant, food, meal, bakery, cafe, meal_delivery, meal_takeaway
home	highest dwell time location from 3 - 5 AM of an individual
work	highest dwell non-home location from 8 AM - 6 PM, 6 PM - 11 PM, 11 PM - 3 AM.

TABLE 3.1: Activity groups

3.3.1 Locations to Activity Trajectories

Prior studies have used sensor data to study association of micro activities, such as daily sleep patterns, gait, and activity rhythms, with health (Saeb et al., 2015; Robben, Pol, and Kröse, 2014). Our mapping of individuals' locations to POI categories, such as restaurants and groceries, opens up a new realm of possibilities to study both macro and micro patterns of an individual. For instance, macro movement and temporal patterns across competing brands inferred from such mapping were used to decide the placement of a new franchise. Further, micro, day-to-day individual-specific patterns such as number of visits, time spent at various business types can predict the individual's next likely location (Molitor et al., 2019).

To identify individual lifestyles, we map the locations to POI categories by using Google Places API³ (second column of Table 3.1) and use the SafeGraph definitions of work to define *home*, *work*, *full-time*, and *part-time* work⁴. Next, we group POI categories with similar semantics (first column in Table 3.1) to

³POIs can be readily identified by matching the longitudes/latitudes using Google Places API https://developers.google.com/places/web-service/supported_types

⁴Social Distancing Metrics Schema by SafeGraph <https://docs.safegraph.com/docs/social-distancing-metrics>

form 15 activity groups that form the universe of all activities a_j^i . Further, to abstract away variations of the exact time in day-to-day activities, a coarser timestamp of t_j^i (timestamp associated with an individual's location), c_j^i is associated with each activity : 12 - 2 AM, 3 - 5 AM, 5 - 7 AM, 7 - 9 AM, 9 - 11 AM, 11 - 2 PM, 2 - 5 PM, 5 - 7 PM, 7 - 9 PM, 9 - 12 PM. The resulting tuples of $d_j^i = (a_j^i, c_j^i)$ across individual trajectories form the universe (W as defined in Def. 8) of temporal activities d_j^i .

3.3.2 Lifestyle Identification

Automatic discovery of individual lifestyles from location data is a non-trivial problem given the massive scale and high dimensionality. Besides, the differences in an individual's activities across days, and the differences from other individuals' activities add further complexity. We take an unsupervised topic modeling approach that has shown potential for uncovering complex temporal and behavioral patterns to identify *work*, *home*, and *consumption* routines (Sun et al., 2014; Farrahi and Gatica-Perez, 2011) on smaller location data sets. Specifically, we leverage the concept of probabilistic Author Topic Model (ATM), designed for text documents (Rosen-Zvi et al., 2012) to model an individual's day-to-day activities. Leveraging the granular location data, we extend this line of literature by incorporating an extensive set of 15 POI or activity types to represent an individual's lifestyle.

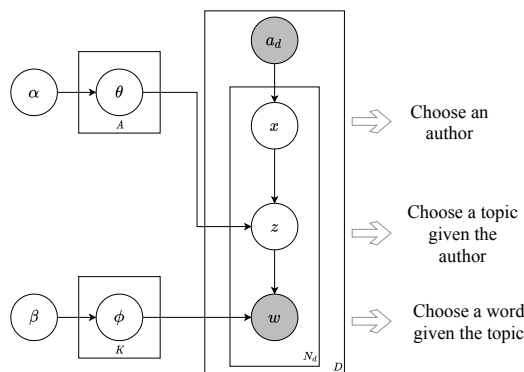


FIGURE 3.1: Probabilistic Graphical model of Author Topic Model using plate notation.

Author Topic Model:

LDA is a probabilistic, unsupervised learning model of a bag of words and of hidden discrete variables called topics. For text modelling, we may view each document as a mixture of various topics, where each topic is characterized as a distribution over words. ATM (Rosen-Zvi et al., 2012) subsumes LDA and assumes authors of documents represent a multinomial distribution over topics where each topic is a probability distribution over words. A document with multiple authors has a distribution over topics that is a mixture of the distributions associated with the authors. When generating a

document, an author is chosen at random for an individual word in the document. This author picks a topic from their multinomial distribution over topics and then samples a word from the multinomial distribution over words associated with that topic. This process is repeated for all words in the document. Formally, the probability of a word w_t assuming K topics, A authors, D documents and W unique words is: $P(w_t) = \sum_{k=1}^K P(w_t|z_t = k)P(z_t = k)$ where z_t is a latent variable showing the topic from which the t^{th} word is drawn. The aim of ATM inference is to determine the word distribution $P(w|z = k) = \phi_w^{(k)}$ for each topic k and the distribution of topics for authors $P(a = k) = \theta_k^{(a)}$ for each author a . $P(\theta)$ is a Dirichlet(α) and $P(\phi)$ is a Dirichlet(β), where α and β are hyper-parameters. Gibbs approximation proposed in (Rosen-Zvi et al., 2012) can be used to estimate these as

$$\phi_k^{(w)} = \frac{n_k^{(w)} + \beta}{n_k^{(\cdot)} + W\beta}; \theta_k^{(a)} = \frac{n_k^{(a)} + \alpha}{n_k^{(a)} + K\alpha} \quad (3.1)$$

where $n_k^{(w)}$ and $n_k^{(a)}$ are the number of times word w and author a have been assigned to topic k , respectively. Similarly, $n_k^{(\cdot)} = \sum_{1:W} n_k^{(w)}$, $n_k^{(a)} = \sum_{1:K} n_k^{(a)}$ are the word-topic and author-topic sum, respectively. Next, we detail our ATM-based lifestyle identification from the individual activity trajectories (D_i).

Activity trajectories to Lifestyles:

To identify lifestyles, we make an analogy between text documents and day-to-day activities, authors, and individuals. We view each activity d_i^j in D_i , the mapped activity trajectory as a word w . We represent each day's activities of an individual (author) as a bag of words – document d . We view activities across multiple days of an individual i as unique documents of an author a . Based on these, we estimate the two ATM model parameters $\phi_k^{(d_i^j)}$, $\theta_k^{(i)}$ using Eq. 3.7 which represents the probability of activity for each topic k , and the probability of topics k for an individual i , respectively. Given these probability distributions, we can rank activities for each topic (i.e., lifestyle) discovered. We can also rank topics for individuals which we view as their primary lifestyles.

We represent each lifestyle as the top Y activities ranked by their relevance (Sievert and Shirley, 2014) – a convex combination of topic-specific probability of each activity (first term in Eq. 3.2) and lift (second term in Eq. 3.2, p_{d_j} is the empirical distribution of activity d_j).

$$r(d_j) = \lambda \log(\phi_k^{d_j}) + (1 - \lambda) \log \frac{\phi_k^{d_j}}{p_{d_j}} \quad (3.2)$$

Next, we assign the most probable topic from the estimated author-topic distribution θ_k^i as the primary lifestyle of an individual. Combining this with the top Y activities ranked by relevance, we can represent an individual i 's

lifestyle as $L_i = \{(d_1^i, d_2^i, \dots, d_Y^i)\}$, $d_j \in W$. This completes the identification of the individual's lifestyle L_i from T_i . We augment these macro representations with other facets of social determinants extracted from location data that capture the micro day-to-day activities, accessibility to various resources, and socio-economics of an individual's neighborhood.

3.3.3 Other Social Determinants

In Table 3.2, we describe different facets of individual social determinants extracted from the location data and the proxies used to indicate an individual's health outcome - hospitalization. To construct these, we glean through the literature on the prediction of health outcomes from medical claims (García-Olmos et al., 2019) or EHR data (Hilton et al., 2020) across disciplines and make necessary adaptations to compute them from the individual location data. These features also form the input and output of our prediction model detailed later.

1) **Lifestyles:** We identify individual weekday and weekend lifestyles from their respective activity trajectories using the above ATM.

2) **Activity:** While lifestyles capture an individual's global routines, the behaviorally rich location data also enable us to capture the day-to-day micro activities. We leverage the transformed activity trajectories D_i (as defined in Def. 8) to compute an individual's daily visit frequencies and dwell time for each of the 15 activity groups a_j^i as additional dynamic, numerical individual features.

3) **Mobility:** Mobility metrics have been shown associated with health outcomes (Saeb et al., 2015). They capture an individual's daily mobility patterns based on the locations visited in T_i , such as the individual's frequency to, time spent at (Pappalardo, Rinzivillo, and Simini, 2016), and distance traveled to a location (Williams et al., 2015). We also compute other richer mobility metrics, such as entropy and radius of gyration (Gonzalez, Hidalgo, and Barabasi, 2008). All these are daily, dynamic, numerical, individual level features.

4) **Accessibility:** Recent studies leveraging medical data also reveal the importance of neighborhood social demographics in predicting patient re-admission and length of stay (Hilton et al., 2020). We hence leverage the transformed activity trajectory (D_i) and compute individual *accessibility* - the closest distance to various resources, such as hospitals, parks, fitness centers, pharmacies, public transport, and work from individual's *home* location. All these are static (time-invariant), numerical, individual level features.

5) **Socio-economics:** Based on the individual's *home* location from the transformed activity trajectories and publicly available Census data⁵, we also compute several census block level socio-economic factors as in (Hilton et al., 2020). These are static and comprise of both categorical (*employment_type - part-time/full-time/nowork*) and numerical features (*population* of individual's census block).

6) **Hospitalization:** To identify if an individual has a future hospitalization,

⁵We obtain the Census Block Group (CBG) level data from SafeGraph: <https://docs.safegraph.com/docs/open-census-data#section-censusedemographic-data>.

we overlay the day-to-day location trajectories on the publicly available location repositories of medical facilities. Specifically, we use the public data sets of hospitals, emergency medical services, and urgent care facilities from Homeland Infrastructure Foundation Level Data (HIFLD)⁶. Based on the overlaid data of medical facilities, we construct proxies to indicate the occurrence of individual's hospitalization event. Specifically, we assign an individual's *hospitalization* = 1 in an observation period, if the individual, whose *work* location is not at a medical facility, has at least 4 hours of activity at a medical facility – two of which occur during late night (12 AM - 5 AM) and the other two during 5 AM - 12 AM. We further assign *hospitalization_night* = 1 if an individual has spent at least two late night hours at a medical facility (12 AM - 5 AM).

3.3.4 Health Risk Quantification

Our quantification of an individual's healthcare risk hinges on learning a model from the location data to accurately predict the future health outcome, hospitalization in our empirical study. We perform both model-free and Logit Regression analyses; and find consistent, qualitative (Figures 3.7b, 3.7d) and quantitative (Table 3.6) evidence that different lifestyles leads to heterogeneous rates of hospitalization.

Modelling Hospitalization

Multiple types of (dynamic, static, categorical, numerical) individual features (Table 3.2) can capture multi-faceted social determinants, but also entail modeling challenges, such as the need to jointly represent all feature types, account for feature interactions, and concatenate features strategically to circumvent a sub-optimal predictive model. We address these challenges by separately learning the representations of the dynamic and static features that account for the interactions among different types of features. Next, we combine these, allowing for the interactions among the two representations, to learn a final joint representation of all the features. To achieve this, we represent the dynamic features by a Context-LSTM (CLSTM) cell proposed by (Ghosh et al., 2016), a modification of the traditional LSTM cell, widely used for word translation and time series modelling. CLSTM incorporates both dynamic and static contextual features to a time series. In (Ghosh et al., 2016), the dynamic contextual features are the latent topics that are jointly represented with words; and each word of the time series is concatenated with an embedding of the topic to predict the next likely word in a sentence. Extending this to our setting, lifestyles (**Lifestyle** features in Table 3.2) are latent topics learned from different activities and serve as a context to the activity-related dynamic features (**Activity** in Table 3.2). Viewing lifestyles as a context to the other dynamic features (**Mobility** in Table 3.2) also leads

⁶The latitudes and longitudes of hospitals, emergency medical services, and urgent care reported by state and federal resources are available at <https://hifld-geoplatform.opendata.arcgis.com/datasets/>

to better predictions⁷. Next, we concatenate these representations for a given time period with the embeddings of the static categorical and numerical features (**Social Demographics** and **Accessibility**) to jointly learn the representation of all features to predict an individual's future hospitalization. Such concatenations of multiple views of an individual's features to form a unified representations are widely studied in multi-modal learning (cf. (Ramachandram and Taylor, 2017) for a review).

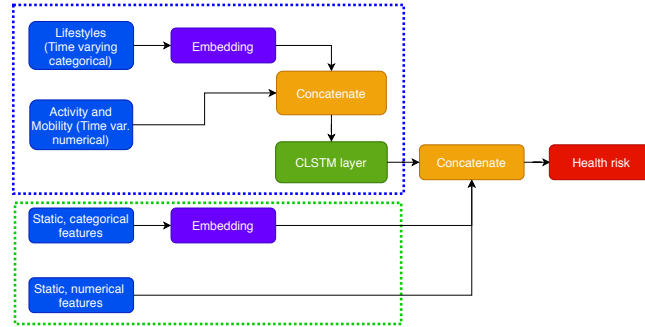


FIGURE 3.2: Architecture diagram of the proposed learner.

An overview of the architecture diagram of the proposed sequential deep learning model is presented in Figure 3.2. The blue box in the figure illustrates the modelling of an individual's temporal features at a day-level with a CLSTM cell (multiple days as a CLSTM layer), where the lifestyle serves as a context for the activity and mobility features. The green box shows the representations of the individual's static features, which are later concatenated with the temporal representations to predict the individual's hospitalization. Next, we formally detail the transformations performed by various layers (non-blue boxes in Figure 3.2) in the proposed learner.

Proposed Learner

Let X_{TN} denote the dynamic numerical feature tensor (number of users \times number of days in the observation period \times number of dynamic numerical features), X_{TC} the dynamic categorical feature tensor (number of users \times number of weeks⁸ \times number of dynamic categorical features), matrices X_{SN} and X_{SC} (number of users \times number of categorical/numerical features) the static numerical, categorical individual features, respectively. To simplify the notation, in the following discussion, we will focus on a single individual's features denoted by \mathbf{x}_{TC} , \mathbf{x}_{TN} , \mathbf{x}_{SC} , and \mathbf{x}_{SN} and their transformation to the probability of future hospitalization (i.e., the health risk).

1) **Embedding:** Embedding layers transforms one-hot encoded categorical features (\mathbf{x}_{TC} , \mathbf{x}_{SC}) to a continuous vector representation of a fixed dimension. Formally,

$$\mathbf{e}_{TC} = \mathbf{x}_{TC} W_{TC}^e; \mathbf{e}_{SC} = \mathbf{x}_{SC} W_{SC}^e \quad (3.3)$$

⁷This is not surprising since an individual's lifestyle is likely correlated with his/her daily mobility behavior and hence a better predictor of his/her health outcome when we explicitly factor in both the lifestyle and mobility behavior.

⁸Lifestyles are the only dynamic categorical features (weekday/weekend). Both dynamic and static categorical features are encoded using a one-hot encoding scheme.

Feature grouping	Name	Definition	Time Varying	Baltimore				D.C.				
				Mean	Std. Dev.	Min	Max	Mean	Std. Dev.	Min	Max	
Lifestyle	lifestyle	Weekend and weekday lifestyle	✓	Refer Figure 3.5, 3.6				Refer Figure B.1, B.2				
	home_freq	Daily frequency & dwell time at individual home	✓	5.99	17.1	1	249	5.93	18.0	1	237	
	home_dwell		✓	1.68	2.80	0	24	1.55	2.82	0	24	
	health_freq	Daily frequency & dwell time at health activity	✓	1.42	7.14	0	183	1.07	6.06	0	232	
	health_dwell		✓	0.43	1.42	0	13.49	0.33	1.13	0	12.17	
	necessityshopping_freq	Daily frequency & dwell time at necessity shopping	✓	1.72	8.07	0	216	1.46	7.29	0	215	
	necessityshopping_dwell		✓	0.59	1.58	0	7.79	0.43	1.30	0	6.83	
	publictransport_freq	Daily frequency & dwell time at public transport	✓	1.26	6.31	0	238	2.35	10.2	0	229	
	publictransport_dwell		✓	0.48	1.34	0	6.85	0.71	1.80	0	6.89	
	religious_freq	Daily frequency & dwell time at religious places	✓	0.73	4.82	0	162	0.591	4.39	0	198	
	religious_dwell		✓	0.26	1.07	0	5.53	0.19	0.875	0	4.39	
	work_freq	Daily frequency & dwell time at work	✓	3.21	8.67	0	233	4.20	10.3	0	192	
	work_dwell		✓	1.32	2.36	0	24	1.26	2.64	0	24	
	Activity	hospital_freq	Daily frequency & dwell time at hospitals	✓	0.13	2.20	0	164	0.08	2.13	0	156
		hospital_dwell		✓	0.04	0.43	0	24	0.02	0.24	0	24
personalcare_freq		Daily frequency & dwell time at personal care	✓	0.30	3.47	0	208	0.281	2.97	0	161	
personalcare_dwell			✓	0.09	0.59	0	2.55	0.08	0.55	0	1.96	
restaurant_freq		Daily frequency & dwell time at restaurants	✓	0.78	4.22	0	145	1.32	5.90	0	177	
restaurant_dwell			✓	0.27	0.83	0	3.92	0.41	1.07	0	4.21	
unhealthyactivities_freq		Daily frequency & dwell time at unhealthy activities	✓	0.16	2.21	0	140	0.15	0.87	0	162	
unhealthyactivities_dwell			✓	0.04	0.39	0	4.32	0.02	0.22	0	12.6	
leisureshopping_freq		Daily frequency & dwell time at leisure shopping	✓	0.26	2.45	0	149	0.28	2.40	0	173	
leisureshopping_dwell			✓	0.10	0.59	0	4.91	0.09	0.54	0	5.88	
hotel_freq		Daily frequency & dwell time at hotels	✓	0.32	2.68	0	122	0.59	3.92	0	143	
hotel_dwell			✓	0.04	0.46	0	24	0.05	0.31	0	24	
owntransport_freq		Daily frequency & dwell time in own transport	✓	0.24	2.89	0	167	0.41	3.77	0	227	
owntransport_dwell			✓	0.12	0.63	0	24	0.11	0.60	0	24	
n_locations		Number locations in a day	✓	22.2	40.4	3	1585	23.1	42.7	3	1807	
avg_distance	Average distance traveled in a day (km.)	✓	7.42	6.70	0	126	7.59	7.54	0	170		
avg_location_entropy	Shannon entropy of frequency of visits	✓	1.90	1.11	0	1	1.84	1.16	0	5.87		
avg_time_entropy	Shannon entropy of dwell time at locations	✓	1.68	1.17	0	5.46	1.58	1.18	0	5.47		
n_unique_locations	Number of unique locations in a day	✓	7.6	15.7	1	573	8.94	18.6	1	417		
avg_time_spent	Average time spent at locations (in hours)	✓	4.20	3.64	0	24	4.08	3.61	0.06	24		
avg_rdg	Average radius of gyration from home (in km.)	✓	6.11	4.28	0	125.1	6.21	4.71	0	132.1		
avg_speed	Average speed during the day (kmph)	✓	6.92	10.57	0	129	6.51	11.80	0	134		
Accessibility	hospital_access	Distance from home to closest hospital (km.)	✗	1.63	0.84	0.02	2.62	1.58	0.89	0.21	4.1	
	park_access	Distance from home to closest park	✗	0.40	0.29	0.04	2.67	1.24	0.21	0.03	4.62	
	fitness_access	Distance to closest fitness facility	✗	0.55	0.37	0.02	2.04	0.76	0.42	0.03	2.91	
	prescription_access	Distance to closest pharmacy	✗	0.45	0.28	0.02	2.05	0.61	1.25	0.02	2.69	
	commute_access	Distance to closest commute	✗	0.15	0.13	0.02	3.05	0.23	0.18	0.02	2.92	
work_access	Distance from home to work	✗	1.90	3.34	0	39.4	1.86	3.37	0	41.1		
Social Demographics	employment_type	Employment type of individual	✗	-	-	-	-	-	-	-	-	
	employment_percent	Employment % in individual's census block group (cbg)	✗	0.82	0.09	0.44	1	0.85	0.09	0.59	1	
	health_ins_percent	Health insurance % in individual's cbg	✗	0.99	0.04	0	1	0.98	0.04	0.07	1	
	population	Population in individual's cbg	✗	1145	561	3	4696	1551	862	8	5254	
	household_income	Average household income in cbg	✗	58321	31951	8654	250000	94193	50226	10278	250000	
	median_age	Median age in individual's cbg	✗	37.4	9.30	10.8	79.9	35.7	7.53	18.9	73.8	
Health Outcome	gross_rent	Gross rent in individual's cbg	✗	240	241	0	1384	395	246	0	2082	
	hospitalization	spent 4 hours in a day at a medical facility (2 during 5AM - 12 AM, 2 during 12 AM - 5 AM)	✗	0.024	0.15	0	1	0.026	0.16	0	1	
hospitalization_night	An indicator if an individual spent two late night hours (12 AM - 5 AM)	✗	0.028	0.16	0	1	0.029	0.17	0	1		

TABLE 3.2: Definition and Summary Statistics of Social Determinants and Health Events

where W_{TC}^e – number of dynamic categorical features $\times N_{TC}^e$, W_{SC}^e – number of static categorical features $\times N_{SC}^e$ are the learnable weight parameters, N_{TC}^e , N_{SC}^e are tune-able model hyper-parameters. Recall that in our setting, X_{TC} comprises of weekday and weekend lifestyles (L_i), both represented by top ten relevant activities (d_j^i , universe of activities W). Hence, an individual's weekday and weekend lifestyle can both be represented as a vectors of length $|W|$; that is, we learn two weight matrices of dimensionality $|W| \times N_{TC}^e$ to compute \mathbf{e}_{TC} . A similar procedure is followed to transform the other static categorical features (*employment_type*).

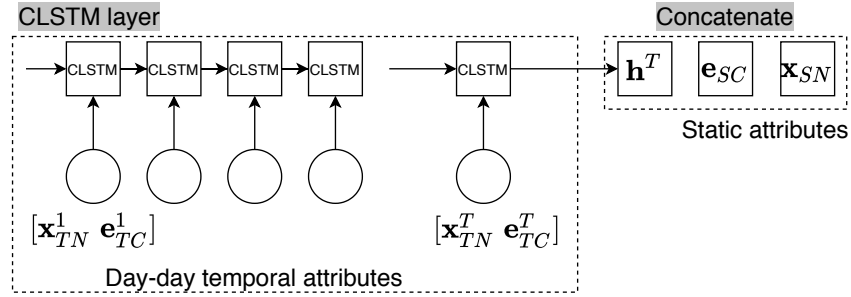


FIGURE 3.3: Illustrations of CLSTM and Concatenate layers.

2) **CLSTM layer:** CLSTM layer (illustrated in Figure 3.3) comprises of multiple CLSTM cells, each of which acts on different days of the $\mathbf{x}_{TN}, \mathbf{e}_{TC}$. Assume $\mathbf{x}_{TN}^t, \mathbf{e}_{TC}^t$ correspond to all numerical, embedded categorical dynamic features (contexts as defined in (Ghosh et al., 2016)) on an arbitrary day⁹, each CLSTM cell performs the following transformations:

$$\begin{aligned}
 i^t &= \sigma(W_{iTC}\mathbf{e}_{TC}^t + W_{iTN}\mathbf{x}_{TN}^t + W_{ih}h^{t-1} + b_i) \\
 f^t &= \sigma(W_{fTC}\mathbf{e}_{TC}^t + W_{fTN}\mathbf{x}_{TN}^t + W_{fh}h^{t-1} + b_f) \\
 o^t &= \sigma(W_{oTC}\mathbf{e}_{TC}^t + W_{oTN}\mathbf{x}_{TN}^t + W_{oh}h^{t-1} + b_o) \\
 c^t &= f^t * c^{t-1} + i^t * \tanh(W_{cTC}\mathbf{e}_{TC}^t + W_{cTN}\mathbf{x}_{TN}^t + W_{ch}h^{t-1} + b_c) \\
 h^t &= o^t * \tanh(c^t)
 \end{aligned} \tag{3.4}$$

The above four equations detail modifications of the traditional LSTM cell where i, f and o are the input, output, and forget gates, respectively, to incorporate additional context \mathbf{e}_{TC}^t . After rearranging the terms, this is equivalent to considering a composite input $[\mathbf{x}_{TN}^t \ \mathbf{e}_{TC}^t]$. since

$$\begin{aligned}
 i^t &= \sigma([W_{iTC} \ W_{iTN} \ W_{ih} \ 1][\mathbf{e}_{TC}^t \ \mathbf{x}_{TN}^t \ h^{t-1} \ b_i]^T) \\
 f^t &= \sigma([W_{fTC} \ W_{fTN} \ W_{fh} \ 1][\mathbf{e}_{TC}^t \ \mathbf{x}_{TN}^t \ h^{t-1} \ b_f]^T) \\
 o^t &= \sigma([W_{oTC} \ W_{oTN} \ W_{oh} \ 1][\mathbf{e}_{TC}^t \ \mathbf{x}_{TN}^t \ h^{t-1} \ b_o]^T) \\
 c^t &= f^t * c^{t-1} + i^t * \tanh([W_{cTC} \ W_{cTN} \ W_{ch} \ 1][\mathbf{e}_{TC}^t \ \mathbf{x}_{TN}^t \ h^{t-1} \ b_c]^T) \\
 h^t &= o^t * \tanh(c^t)
 \end{aligned} \tag{3.5}$$

⁹ \mathbf{e}_{TC}^t is computed depending on whether the day is a weekday or weekend, since our lifestyles are derived for weekday/weekend rather than days.

Each CLSTM cell transforms the concatenated input $[\mathbf{e}_{TC}^t \mathbf{x}_{TN}^t]$ into a hidden representation h^t (dimensions : number of individuals $\times N_T^e$) with learnable shared¹⁰ weight and bias parameters (W_*, b_*) and tune-able hyper-parameter N_T^e . Hence, the resulting representations from the CLSTM layer are $\{h^t\}, t \in [1, T]$, where T is the number of days in our observation period.

3) **Concatenate:** Concatenate layers do not contain any learnable parameters and are simply used to combine different intermediate representations. We perform two concatenations, $[\mathbf{x}_{TN}^t \mathbf{e}_{TC}^t]$ as illustrated in Figure 3.3). Second, the concatenation of the hidden temporal representation obtained from the CLSTM layer ($\{h^t\}$), embedded static (\mathbf{e}_{SC}) and numerical features (\mathbf{x}_{SN}). Noting that \mathbf{h}^T , the hidden layer representation of the last day of observation captures temporal relations across the preceding days due to the recurrence nature of Equations 3.5, we combine this with $\mathbf{e}_{SC}, \mathbf{x}_{SN}$ to form $[\mathbf{h}^T \mathbf{e}_{SC} \mathbf{x}_{SN}]$, the final joint representation which comprises of both the dynamic and static features.

4) **Health risk:** We pass on the final representation into a fully connected dense layer, allowing for interactions between the temporal and static features, and assign the probability of hospitalization as

$$r = \sigma([W_T W_{SC} W_{SN} 1][\mathbf{h}^T \mathbf{e}_{SC} \mathbf{x}_{SN} b_r]^T) \quad (3.6)$$

where W_T, W_{SC}, W_{SN}, b_r are learnable parameters. For a given binary health outcome (hospitalization), to learn the various weights (W_* in Equations 3.3, 3.5, 3.6), we minimize the binary cross-entropy loss between the observed health outcome and \mathbf{r} , the vector of outcome probabilities. The rest of the hyper-parameters are tuned via cross-validation (details in Section 3.5).

3.4 Data

We combine several data sets: individual-level smartphone location data, census-block-level demographic data from the American Community Survey (2016), and HILFD public data of hospitals, emergency medical services, and urgent care facilities. The location data are curated with privacy compliance by a leading data collector via hundreds of commonly used mobile apps. The data cover one-quarter of the U.S. population across Android and iOS operating systems. Each data record corresponds to a location tracked with information about 1) Individual ID: an anonymized unique identifier of an individual's device, 2) Latitude, longitude and timestamp of a location visited; 3) Speed at which a visit was captured.

We analyzed data samples from Baltimore and D.C. (Baltimore – October, November 2018 and 2019; D.C. – April, May 2018 and October, November 2019). For each city, we only analyze the individuals who appear across all four months and at least ten days per month. We also eliminate those without reliable identification of *work* and *home* locations. The final sample comprises of 4,528 from Baltimore and 6,114 individuals from D.C. Tables 3.2, 3.3 and Figure 3.4 display the summary statistics of the social determinants of health

¹⁰All the learnable weights W_* and bias parameters b_* are shared across different time steps (days in our model).

Description	Baltimore				D.C.			
	Weekday		Weekend		Weekday		Weekend	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Daily activities	14.64	17.12	13.81	19.64	14.70	18.48	15.86	19.66
Unique daily activities	10.71	10.42	9.62	11.12	10.81	11.38	11.34	11.70
Activities at <i>home</i>	5.15	14.3	8.89	22.2	4.86	14.5	8.32	21.9
Activities at <i>work</i>	4.37	9.43	0.79	4.65	4.89	11.2	0.95	5.85
Activities at <i>publictransport</i>	2.25	6.03	2.30	7.20	2.29	9.94	2.57	10.9
Activities at <i>other</i>	4.86	15.4	5.13	17.2	4.73	15.2	4.99	16.4

TABLE 3.3: Summary statistics of the activity trajectories

computed from the location data, census block demographics, and public medical facilities.

Location Trajectories: Table 3.2 (**Mobility** row) shows the summary statistics of the raw location data. In Baltimore, there are on average 22 total locations (and 7 unique locations) per individual per day. The average speed is 6.92 kmph. For all other measures, we eliminate the locations captured at a speed > 5 kmph and dwell time < 5 minutes¹¹. The average Haversine Distance between consecutive locations is 7.42 km and the average dwell time 2.2 hours.

Activity Trajectories : Table 3.2 (**Activity** and **Accessibility** rows), Table 3.3 detail the summary statistics of the activity trajectories (Section 3.3.1 for the transformation of the locations to activity trajectories). Table 3.2 shows that out of the 15 activity groups (Table 3.1), *home*, *work* and *publictransport* are the top three in both the average daily occurrences and time spent. When broken down by weekday and weekend (Table 3.3, Baltimore), *work* occurs less frequently (0.79) during the weekend compared to weekdays (4.37). In contrast, *home* occurs more frequently during weekends (5.15) than weekdays (8.89). To accommodate the differences¹² in these top activities, we learn weekday and weekend lifestyles separately. There are on average 14.64 total daily activities (9.62 unique) per individual on weekdays, characterized by its activity group and time range (*restaurant* (2 - 5 PM)); and 14 (10.71 unique) on weekends.

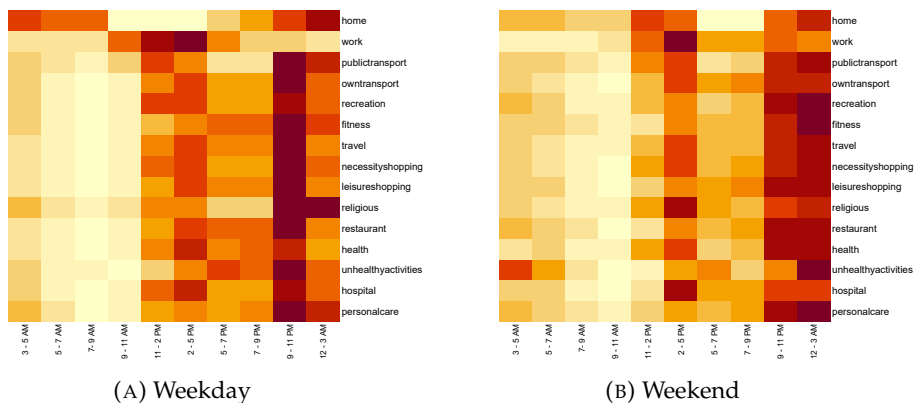


FIGURE 3.4: Row-normalized heatmap of activity occurrences (Baltimore). Darker reds indicate higher occurrences.

¹¹The time difference between consecutive locations is used to determine the dwell time spent at a location.

¹²We confirmed that these differences between weekends and weekdays for *work* and *home* are statistically significant at $p = 0.01$ based on a paired Wilcoxon test.

Figure 3.4 plots the heat map of the activities on weekends and weekdays, with a lighter color indicating a lower occurrence of an activity during the corresponding time. Figure 3.4a shows that *work* mostly occurs during 2 - 5 P.M., *home* 12 - 3 AM. In contrast, on weekends people tend to stay *home* during the same time window (Figure 3.4b). Also, leisure, shopping, and consumption activities, occur earlier (9 - 11 P.M.) on weekdays than weekends (12 - 3 A.M.).

Census Block Socio-economics: An individual's census block is assigned based on the closest census block by Haversine Distance to his/her *home*. Table 3.2 (**Social Demographics**) exhibits the summary statistics of the census block socio-economic factors.

Health Outcome: An individual's health outcome is defined as a hospitalization event observed in the location data over the last two months of the sampling period. A total of 111 (158) individuals had hospitalizations spanning both day and night and 127 (175) during night at Baltimore (D.C.).

3.5 Empirical Study

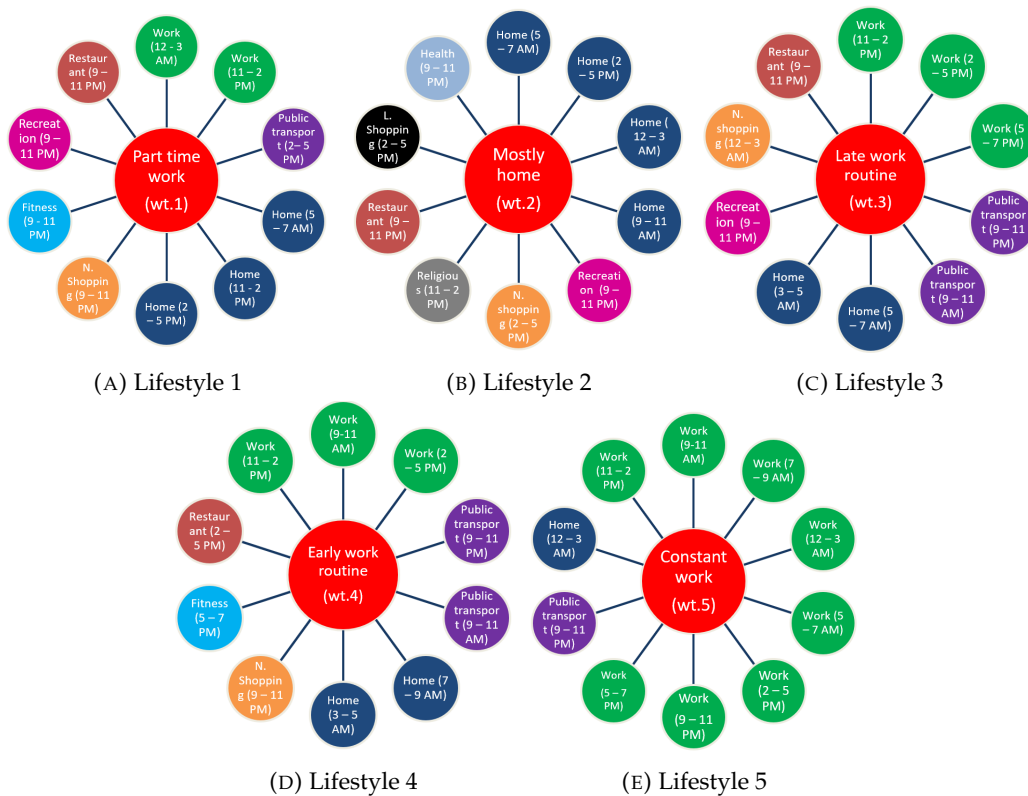


FIGURE 3.5: Weekday Lifestyles (Baltimore)

3.5.1 Lifestyles

We apply the proposed ATM-based methodology on the first two months of the location data during the sampling period to identify the lifestyles. Figures 3.5 and 3.6 present the top 10 activities representative of weekend and

weekday lifestyles. The number of lifestyles (topics, K) are decided based on coherence (Sievert and Shirley, 2014). Coherence measures how well-focused the top words (activities) describe a specific lifestyle. We vary K between 3 to 25, and compute the average coherence over 50 runs to determine the number of topics for weekdays and weekends, respectively. The top 10 ($Y = 10$, Def. 9) relevant activities for each topic are then visualized for the highest coherent ATM model in figures 3.5 and 3.6. In total, we identify five weekday and four weekend lifestyles with different activities across different hours-of-the-day.

Weekday Lifestyles: Figure 3.5 visualizes the five identified weekday lifestyles and their corresponding activities for Baltimore residents. Lifestyle 3 (denoted by $wt.3$) characterizes a late work routine ($work$ over 11 - 2 PM, 2 - 5 PM, 5 - 7 PM), commute via public transportation mornings and evenings ($publictransport$ over 9 - 11 AM, 9 - 11 PM), late night dining at restaurants, grocery shopping, and recreation ($necessityshopping$ 12 - 3 AM, $restaurant$ 9 - 11 PM, $recreation$ in 9 - 11 PM). In contrast, lifestyle 4 ($wt.4$, although with similar commute and consumption patterns, reveals an early work routine: $work$ 9 - 11 AM, 11 - 2 PM, 2 - 5 PM, and fitness during evenings (5 - 7 PM). Both lifestyles feature a steady full-time work routine, and work-fitness balance. Lifestyle $wt.1$, in comparison, indicates a part-time job ($work$ 11 - 2 PM, 12 - 3 AM).

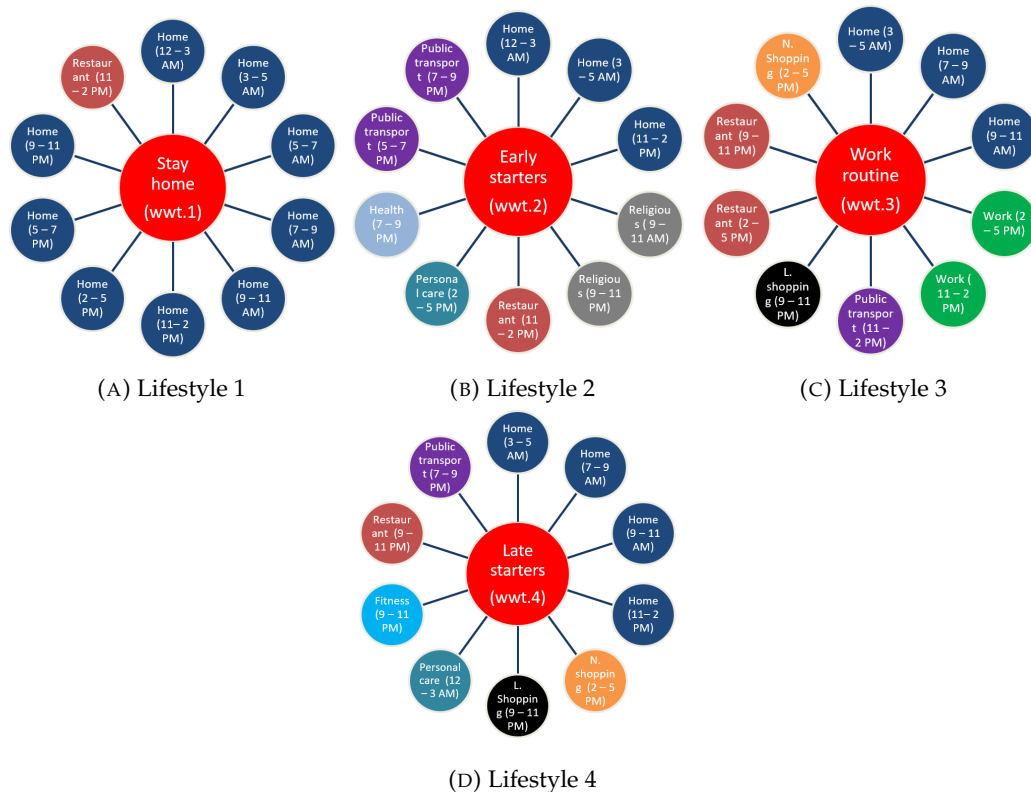


FIGURE 3.6: Weekend Lifestyles (Baltimore)

Weekend Lifestyles: Figure 3.6 displays the top ten activities of the four weekend lifestyles for Baltimore residents. As expected, apart from lifestyle $wwt.3$ ($work$ 11 - 2 PM, 2 - 5 PM), all other lifestyles suggest a non-work routine. Lifestyle $wwt.1$ characterizes an early start weekend routine with visits

to religious locations (*religious* 9 - 11 AM, 9 - 11 PM) and *restaurant* afterwards (11 - 2 PM). In contrast, lifestyle *wwt.4* indicates a late start routine, where the individuals mainly stay at *home* during these hours, with fitness and recreations later in the evening (*fitness* 9 - 11 PM, *personalcare* 12 - 3 AM). Besides work on weekends, individuals in lifestyle *wwt.3* regularly consume at restaurants (*restaurant* 2 - 5 PM, 9 - 11 PM). Lifestyle *wwt.1* stays at home weekends, dine at *restaurant* 11 - 2 PM, with limited fitness or leisure activities. In D.C., we identify five weekday and four weekend lifestyles. Due to space limitations, we focus on Baltimore. Overall, the proposed lifestyle identification uncovers distinctive activity patterns from location data.

3.5.2 Health Risk Quantification

We identify hospitalization from the two months of location data in 2019 and then link them to the social determinants.

Model-free Evidence: Figure 3.7 exhibits the histogram of the percentage of the 4,528 Baltimore residents with each lifestyle visiting medical facilities. Weekday lifestyles *wt.2* and *wt.5* have higher (3.29% and 4.95%, Figure 3.7a) than average (2.45%) percentages of individuals visiting medical facilities. In contrast, lifestyle *wt.4* has about half (1.49 %) the average percentage of hospitalizations. Similarly, Figure 3.7c reveals that weekend lifestyles *wwt.1* and *wwt.3* experience higher percentages of hospitalizations whereas lifestyle *wwt.2* half less likely.

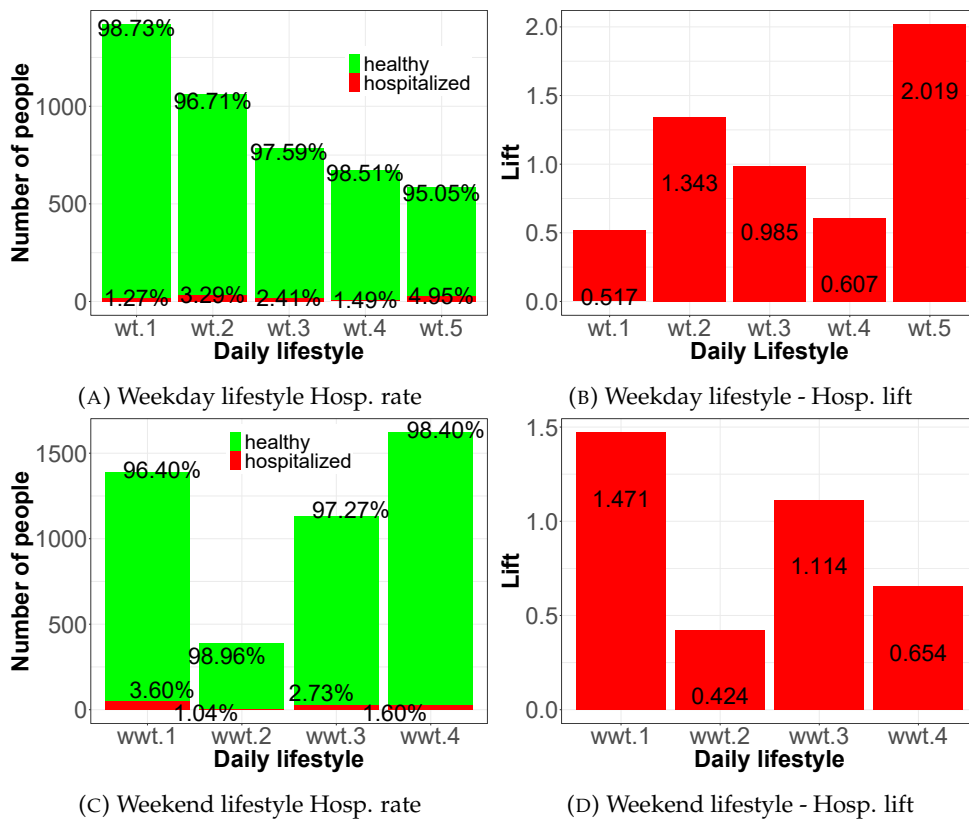


FIGURE 3.7: (Baltimore) Association with Hospitalization : Model free analysis (*hospitalization*)

We quantify the relative rate of future hospitalizations per lifestyle against the average rate by using the lift score (Figures 3.7b, 3.7d). The top activities characterizing each lifestyle (Figures 3.5, 3.6) and their lift scores suggest that those who participate in *fitness* on weekends (*wwt.4*) or weekdays (*wt.4*) are less likely (0.65 and 0.60, respectively) to have hospitalizations on average. On the other extreme, those with either busy, volatile work routines (*wt.5*) or no work routine (*wt.2*) on weekdays, are 2.01 and 1.34 times more likely to have hospitalizations (Figures 3.7b, 3.5; and people who either work (*wwt.3*) or mostly stay at home (*wwt.1*) on weekends are 1.1 to 1.4 times more likely to have hospitalizations than average (Figures 3.7d, 3.6). Overall, the model-free evidence reveals heterogeneous rates of hospitalizations across different lifestyles.

Logit Analysis: To supplement the model-free evidence, we examine an individual's likelihood of having future hospitalization using a Logit model:

$$P(\text{hospitalization}_i) = \frac{e^{X_i}}{1 + e^{X_i}} \quad (3.7)$$

$$X_i = \alpha_i + \beta_1 \text{lifestyle_weekday}_i + \beta_2 \text{lifestyle_weekend}_i + \beta_3 X_i^{\text{access}} + \beta_4 X_i^{\text{mobility}} + \beta_5 X_i^{\text{demog}} + \beta_6 X_i^{\text{community}}$$

where *hospitalization_i* is 1 if an individual had at least one hospitalization during the two months in 2019, *lifestyle_weekday_i* and *lifestyle_weekend_i* are dummies indicating the individual's weekend and weekday lifestyles, X_i^{access} , X_i^{mobility} are the average daily accessibility and mobility metrics, respectively, in Table 3.2; $X_i^{\text{community}}$ are the average mobility and accessibility metrics of the residents in the same census block group as the individual; and X_i^{demog} are the census block socio-economic factors. Table 3.4 (Columns 1 - 4) displays the maximum log-likelihood estimates of the lifestyles while controlling for different individual-level features. The coefficients indicate the odds of an individual with a specific lifestyle to have a future hospitalization over the average odds.

	Dep. variable : hospitalization				
	(1)	(2)	(3)	(4)	(5)
Weekend lifestyle 1 (wwt.1)	0.384** (0.174)	0.353** (0.176)	0.324** (0.182)	0.296* (0.197)	0.295* (0.198)
Weekend lifestyle 2 (wwt.2)	-0.804** (0.347)	-0.771** (0.348)	-0.805** (0.354)	-0.976** (0.402)	-0.982** (0.404)
Weekend lifestyle 4 (wwt.4)	-0.024 (0.182)	-0.010 (0.183)	0.114 (0.192)	0.213 (0.208)	0.214 (0.210)
Weekday lifestyle 1 (wt.1)	-0.685*** (0.189)	-0.691*** (0.190)	-0.665*** (0.200)	-0.654*** (0.209)	-0.648*** (0.211)
Weekday lifestyle 2 (wt.2)	0.121 (0.177)	0.076 (0.182)	-0.204 (0.175)	-0.138 (0.193)	-0.130 (0.195)
Weekday lifestyle 4 (wt.4)	-0.494** (0.241)	-0.463* (0.243)	-0.429* (0.250)	-0.395* (0.261)	-0.334* (0.262)
Weekday lifestyle 5 (wt.5)	0.998*** (0.146)	1.030*** (0.149)	0.933*** (0.155)	0.946*** (0.164)	0.928*** (0.167)
Accessibility metrics	✓	✓	✓	✓	✓
Mobility metrics	✓	✓	✓	✓	✓
Social Demographics	✓	✓	✓	✓	✓
Community Controls	✓	✓	✓	✓	✓
Observations	4,528	4,528	4,528	4,528	4,528
Log Likelihood	-665.904	-657.063	-586.244	-532.423	-527.367

*p<0.1; **p<0.05; ***p<0.01

TABLE 3.4: (Baltimore) Hospitalization Logit Analysis

Table 3.4 (Column 4) indicate that those with *wwt.4*, *wt.5* have significantly higher odds of having a future hospitalization ($1.34 \approx \exp(0.296)$) and

2.57, respectively) than average, after controlling for other social determinants. Similarly, lifestyles *wwt.2*, *wt.1* and *wt.4* have significantly lower odds than average. These insights are qualitatively consistent with the model free evidence. Interestingly, we do not find any significant association between X_i^{access} , X_i^{demog} and future hospitalizations indicating that two individuals who live in the same neighborhood with similar social demographics, access to parks/fitness facilities, but with different lifestyles, will have different health risks. In Table 3.5, we introduce total dwell time at healthy (fitness, personal care) and unhealthy activities into the regression and observe that *regularity* of healthy activities matters (lifestyle *wt.1*, *wt.4*, *wwt.4*), instead of the total dwell time (e.g., two individuals with similar fitness/work hours per week, but different distribution of these activities across days may lead to different health risks) further highlighting the importance of mining the lifestyle patterns to quantify health risk. On the flip side, we also observe that total time spent at unhealthy activities is significantly correlated to future hospitalization. In D.C, the qualitative findings remain similar. In addition, we find that *regularity* of unhealthy activities associate with significantly higher odds (1.6) of future hospitalization.

	Dep. variable : hospitalization		
	(1)	(2)	(3)
Weekend lifestyle 1 (wwt.1)	0.295* (0.198)	0.295* (0.198)	0.295* (0.199)
Weekend lifestyle 2 (wwt.2)	-0.978** (0.404)	-1.010** (0.407)	-1.002** (0.406)
Weekend lifestyle 4 (wwt.4)	0.208 (0.210)	0.211 (0.210)	0.235 (0.211)
Weekday lifestyle 1 (wt.1)	-0.638*** (0.212)	-0.619*** (0.214)	-0.589*** (0.212)
Weekday lifestyle 2 (wt.2)	-0.126 (0.195)	-0.143 (0.196)	-0.143 (0.197)
Weekday lifestyle 4 (wt.4)	-0.334* (0.262)	-0.329* (0.262)	-0.339* (0.262)
Weekday lifestyle 5 (wt.5)	0.929*** (0.167)	0.915*** (0.167)	0.917*** (0.167)
total_fitness_dwell	-0.001 (0.009)		
total_personalcare_dwell		-0.001 (0.002)	
total_unhealthyactivities_dwell			0.003** (0.001)
Other social determinants	✓	✓	✓

*p<0.1; **p<0.05; ***p<0.01

TABLE 3.5: (Baltimore) Hospitalization : Additional Logit Analysis

Predictive Performance: As detailed in Sec 3.3.4, the proposed learner

City = Baltimore	Day and Night Hospitalization (hospitalization)		Late Night Hospitalization (hospitalization_night)		6-hour hospital visit (hospitalization_alt)	
Hospitalization rate	2.45%		2.80%		4.75%	
Model /Measure	PR AUC	AUC	PR AUC	AUC	PR AUC	AUC
RF (NLI & NAC)	0.05 (1.13 %)	0.63 (1.91 %)	0.06 (1.05 %)	0.61 (2.15%)	0.11 (1.76%)	0.65 (2.04%)
GB (NLI & NAC)	0.06 (1.14%)	0.64 (2.12%)	0.06 (1.32%)	0.63 (1.69%)	0.12 (1.10%)	0.65 (2.31%)
Lasso (ALLAGG)	0.14 (2.27 %)	0.73 (4.53%)	0.13 (2.05 %)	0.72 (4.70 %)	0.21 (2.02 %)	0.70 (4.58 %)
RF (ALAGG)	0.22 (2.54 %)	0.78 (4.06%)	0.21 (2.63%)	0.74 (4.90%)	0.29 (2.53%)	0.74 (4.69%)
GB (ALLAGG)	0.23 (2.47 %)	0.76 (4.06%)	0.21 (2.69%)	0.73 (4.64%)	0.27 (2.73%)	0.71 (4.44%)
LSTM (NL & NAC)	0.15 (1.97%)	0.72 (1.60%)	0.17 (1.39%)	0.74 (2.12%)	0.21 (1.22%)	0.72 (2.86%)
LSTM (NLI)	0.24 (1.05%)	0.79 (2.95%)	0.24 (1.26%)	0.76 (3.46%)	0.38 (1.71%)	0.80 (3.90%)
Full model	0.28 (1.53%)	0.85 (3.67%)	0.29 (1.17%)	0.84 (3.95%)	0.42 (1.47%)	0.86 (3.67%)

TABLE 3.6: (Baltimore) Hospitalization prediction

City = DC	Day and Night Hospitalization (<i>hospitalization</i>)		Late Night Hospitalization (<i>hospitalization_night</i>)		6-hour Hospital visit (<i>hospitalization_alt</i>)	
Hospitalization rate	2.58%		2.87%		5.71%	
Model/Measure	PR AUC	AUC	PR AUC	AUC	PR AUC	AUC
RF (NLI & NAC)	0.07 (1.01 %)	0.63 (2.11 %)	0.06 (0.96 %)	0.65 (1.85%)	0.12 (1.89%)	0.66 (2.11%)
GB (NLI & NAC)	0.06 (1.08%)	0.62 (2.04%)	0.07 (1.14%)	0.66 (1.96%)	0.11 (1.61%)	0.68 (2.41%)
Lasso (ALLAGG)	0.17 (2.75 %)	0.76 (4.13 %)	0.16 (2.84 %)	0.76 (4.96 %)	0.24 (2.85 %)	0.78 (4.97 %)
RF (ALLAGG)	0.23 (2.96 %)	0.80 (4.82%)	0.22 (2.69%)	0.80 (4.90%)	0.30 (2.57%)	0.81 (4.98%)
GB (ALLAGG)	0.24 (2.07 %)	0.79 (4.27%)	0.22 (2.27%)	0.81 (4.64%)	0.31 (2.63%)	0.82 (4.82%)
LSTM (NLI & NAC)	0.16 (1.24%)	0.74 (1.99%)	0.17 (1.91%)	0.75 (2.62%)	0.26 (1.75%)	0.77 (2.72%)
LSTM (NLI)	0.23 (1.97%)	0.81 (2.71%)	0.21 (2.01%)	0.80 (3.22%)	0.35 (1.75%)	0.81 (3.54%)
Full model	0.30 (1.42%)	0.87 (3.12%)	0.32 (1.39%)	0.89 (3.41%)	0.44 (1.87%)	0.90 (3.80%)

TABLE 3.7: (D.C.) Hospitalization Prediction

takes the individual features extracted from the location data in 2018 to predict the health risk (Eq 3.6), i.e., the probability of an individual having a hospitalization in 2019. In practice, given a series of risk scores, a domain expert would ideally set the minimum threshold to deem if an individual has surpasses an "at-risk" threshold. Hence, in Table 3.6, 3.7, we report the average cross-validated PRAUC and ROCAUC and corresponding standard deviation percentages to sweep all possible thresholds¹³.

We compare our learner’s predictive performance with several baselines’ to investigate 1) importance of jointly representing multiple facets of an individual using a sequential model; 2) performance lift provided by individual behavioral features – lifestyles and day-to-day activities. To support 1), we employ non-sequential learners, Random Forest (RF), regularized logistic regression (LASSO), and Gradient Boosting (GB) with aggregated static, dynamic numerical and categorical features (ALLAGG) (Table 3.2)¹⁴. To support 2), we design ablations of the proposed learner and baselines without the dynamic lifestyles (NLI) and activity features (NAC).

Table 3.6 suggest that the proposed learner outperforms both types of baselines considered. The proposed model for health risk quantification has an AUPRC of 0.28 and AUC of 0.85. The best performing non-sequential model performs worse than the proposed learner (0.23 compared to 0.28), indicating the importance of modelling temporal correlations across features via LSTMs. Ablations of the non-sequential models and the proposed learner suggest lifestyles and day-to-day activities, in aggregate and dynamic form, provide a performance lift. The best performing non-sequential model (GB) has a PR AUC increment from 0.06 to 0.23; sequential models from 0.15 to 0.28. Finally, the ablation of the proposed learner without the CLSTM cell performs worse (0.24 PR AUC) than the full model (16 % increase in PR

¹³We also include *hospitalization_alt*, an indicator of an individual spending 6 hours in a medical facility on any day in the 2 months in 2019, in addition to the 2 indicators in Table 3.2. The data are split into 70% training, 15% validation, and 15% test sets; and a ten-fold cross validation is performed. We perform a grid search to optimize several tune-able model hyper-parameters: dimensionality of the dynamic and static categorical embeddings, class weights, learning rate, number and size of various hidden layers.

¹⁴The dynamic features for e.g. daily unique locations are aggregated across days as the average daily unique locations. The categorical features are encoded as one-hot dummies. Several parameters of the baselines are optimized by performing a grid search. We report the average ten-fold cross-validated PR AUC and AUC. SMOTE (Chawla et al., 2002) is used to account for the class imbalance for the non-sequential baselines.

AUC), indicating the importance of lifestyles as the contexts to the dynamic features. These results remain qualitatively similar for the D.C. residents (Table 3.7, 30% increase from ablations).

3.6 Conclusion

We develop a framework to identify individual social determinants of health and quantify their impact on future hospitalizations from granular smartphone location data. Specifically, building on topic models, we first identify an individual's lifestyles; then supplement them with additional accessibility and socio-economic features; through an array of analyses, we quantify the strong connection between the social determinants, particularly lifestyles, and future hospitalization by leveraging sequential deep learning models. This research broadens the prior literature by exploring novel, extensive, behavior-rich data beyond the EHRs at a population scale, thus offering generalizable insights to guide policy making, promote public health, and mitigate the rocketing healthcare costs.

Appendix A

Personalized and Interpretable Privacy Preservation

A.1 Objective Function Analysis

In our proposed method, we frame the problem of preserving privacy at a consumer level. As defined in RQ1 and RQ2, r_i is the consumer's privacy risk of sharing their trajectory data with an advertiser and u_i be the advertiser's benefit of acquiring the trajectory data. Our obfuscation scheme based on suppression of locations has two consumer specific parameters z_i (number of locations to be suppressed) and \vec{s}_i (identity of locations to be suppressed). To maintain the utility-risk trade-off, the data collector's decision would be to find the tuple $\{\vec{s}_i, z_i\}$ to maximize the expected data utility for advertisers, $E(u_i)$ and minimize the consumer privacy risk $E(r_i)$. Let $E^*(u_i)$ and $E^*(r_i)$ denote the expected utility and risk respectively when there is no obfuscation performed on the consumer trajectories. We frame the problem as minimizing the relative decrease in utility $\frac{(E^*(u_i) - E(u_i))}{E^*(u_i)}$ and maximizing the relative decrease in risk $\frac{(E^*(r_i) - E(r_i))}{E^*(r_i)}$. Formally, the objective function that a data collector faces can be written as

$$O(\{\vec{s}_i, z_i\}) = \text{Min}\left(\frac{(E^*(u_i) - E(u_i))}{E^*(u_i)} - \frac{(E^*(r_i) - E(r_i))}{E^*(r_i)}\right) \quad (\text{A.1})$$

Note that, $E^*(u_i)$ and $E^*(r_i)$ vary depending on the type of privacy risk and utility and can be computed a priori as detailed in Section 1.4.3. This results in

$$O(\{\vec{s}_i, z_i\}) = \text{Min}\left(\frac{E(r_i)}{E^*(r_i)} - \frac{E(u_i)}{E^*(u_i)}\right) \quad (\text{A.2})$$

We solve this objective function empirically by varying the parameter $z_i = r_i p$, with a grid of $pack \in [0, 1]$ (Eq 6, Section 1.4.3). This ensures reduction in expected risk - as we increase p , $\frac{E(r_i)}{E^*(r_i)}$ decreases. For instance, when $p = 0$, $E(r_i) = E^*(r_i)$, a positive constant and for $r_i=1$ (high-risk consumer) and $p=1$, $E(r_i)=0$ since there are no locations shared with the advertiser. The grid-based search offers the following benefits for a data collector, compared to an analytical approach.

1. **Flexibility** : A data collector can plug in different quantifications of $E(r_i)$ and $E(u_i)$, pertaining to different privacy risks and advertiser utilities. We illustrate this by examining two key types of privacy risks

and two advertiser applications in the revised version. If a data collector has access to other heuristics to quantify these, for example, $E(r_i)$ can be simply quantified as unique locations of an individual, higher the number, lower the risk, the grid-based approach would still enable them to find an optimal trade-off. This generalizability is lost when we constrain $E(r_i)$ and $E(u_i)$ to a specific functional form which we see as a key contribution to our study. A recent work published in TKDE, Yang et. al. 2018, take this approach where they aim to reduce the inference attack while maintaining the POI based recommendation utility.

2. **Convergence** : If we are to consider a specific functional form of $E(r_i)$ and $E(u_i)$, to solve the minimization using a descent-based approach, one would need to further constrain the class of functions that $E(r_i)$ and $E(u_i)$ belong to. While optimizing difference of convex functions (Refer to Bačák 2011 for a review) is well studied, $E(u_i)$, which we quantify as MAP@k, MAR@k in both our utility functions (POI and Activity prediction) is not convex (Kar et. al. 2015). This results in a difference of convex and non-convex function, which is currently an active area of research in the optimization community and not the key focus of our work. Finally, if we were to just assume that both $E(r_i)$ and $E(u_i)$ are differentiable, analytical computations of the gradients with respect to the parameters $\{\vec{s}_i, z_i\}$ are computationally intensive given the nature of the heuristics involved in quantifying $E(r_i)$ and $E(u_i)$.

A.2 Early Stopping

An exhaustive grid-based approach comes with its own shortcomings. Specifically, the discretization of the grid p , would decide the best trade-off achieved. While considering a finer discretization of p , can remedy this issue, we would run into computational issues in estimating $E(r_i)$ and $E(u_i)$. We partially address the computational issue in our implementation of the proposed obfuscation scheme, by estimating $E(r_i)$ and $E(u_i)$ for different values of $p \in \{0, 0.1, \dots, 1\}$, in parallel. However, exhaustively searching a finer grid would require constraining the search space to remain computationally efficient. To alleviate this, we propose an early-stopping heuristic which improves the current grid-based search by starting from coarser grid intervals of p instead of a fixed grid of points, iteratively estimates $E(r_i)$ and $E(u_i)$ for a finer grid of values efficiently, guided by an acceptable decrease in advertiser's utility. We discuss the algorithm next.

Input: N consumer trajectories $\{T_i\}$, An estimator for $E(r_i)$, $E(u_i)$, where $r_i = \text{PR}(T_i; \{\vec{s}_i, z_i\})$, $u_i = \mathbf{U}(T_i; \{\vec{s}_i, z_i\})$, optionally, an acceptable relative decrease in Advertiser utility U^{acc}

Output: Obfuscated consumer trajectories $\{P(T_i)\}$

1. Start with a coarse set of grid intervals for $G_p \in \{[0, 0.1], [0.1, 0.2], \dots, [0.9, 1]\}$ and possible set of pre-computed \vec{s}_i based on frequency, time spent and recency of locations in T_i .
2. Set $G_{prune} = \phi$.

- (a) **Estimation** : In parallel, repeatedly sample N_s consumers from N
- i. In each iteration, for each \vec{s}_i , sample a p in each grid $g \in G_p$
 - ii. Compute $E(r_i)$ and $E(u_i)$ over M iterations by suppressing the locations using Eq 1.6). The average of M iterations corresponds to the estimates for a specific grid in G_p .
- (b) **Pruning** :
- i. If $\frac{(E^*(u_i) - E(u_i))}{E^*(u_i)} < U^{acc}$, add the corresponding g to G_{prune}
 - ii. In G_{prune} , keep top $\lceil \frac{|G_{prune}|}{2} \rceil$ grids based on the increasingly sorted $\frac{(E^*(r_i) - E(r_i))}{E^*(r_i)}$
 - iii. **Stopping criterion** : If the M paired estimates of $\frac{(E^*(r_i) - E(r_i))}{E^*(r_i)}$ of the top two grid intervals do not have a statistically significant difference under paired t-test statistic or if G_{prune} is empty, pick the obfuscation parameters with highest $\frac{(E^*(r_i) - E(r_i))}{E^*(r_i)}$ in Step 2 a) ii), obfuscate $\{T_i\}$ and return $\{P(T_i)\}$.
- (c) **Candidate Set** :
- i. Construct finer grids for each $g \in G_{prune}$ by splitting each g at their mid-point resulting in a maximum of $|G_p| + 1$ candidate sets.
 - ii. Set $G_p = G_{prune}$ and go to 2.2.

The outline of our early Stopping heuristic is detailed above. The algorithm starts off with a coarse-grained set of grid intervals G_p instead of a fixed set of points. We instantiate $|G_p|$ independent parallel threads at Step 2) similar to the parallel computation in the proposed fixed grid approach. Each thread is responsible to compute the estimates of $E(u_i)$ and $E(r_i)$ for a grid interval g in G_p , across M repeated sample trajectories of size N_s , which are again executed in parallel. In the fixed grid approach, we compute this estimate by averaging across twenty trials for a fixed value of p (Section 5.1 and 5.2) on all the consumer trajectories N .

Next, each child thread involves sampling a value of p in the grid interval (E.g.: 0.21 in grid [0.2, 0.3]), and computing $E(u_i)$ and $E(r_i)$ across the three specifications of \vec{s}_i . The average of the M resulting estimates for each g is used to prune G_p , to remain computationally efficient and generate finer grid intervals in the successive iterations, thus performing an exhaustive search of the parameter space. In Step 2.2) i), an optional parameter – acceptable relative decrease in advertiser’s utility U^{acc} is used to prune grid intervals in G_p into G_{prune} . We further prune G_p by dropping the grid intervals corresponding to the bottom quantile of the relative decreases in consumer risk $\frac{(E^*(r_i) - E(r_i))}{E^*(r_i)}$. If the resulting G_{prune} is empty, or if the means of the top two estimates in G_{prune} are not statistically significant under paired t-test statistic, we stop the search and obfuscate $\{T_i\}$ based on the parameters that resulted in the highest relative decrease in consumer risk in Step 2.1) ii). Next, we generate a finer candidate set based on the resulting non-empty G_{prune} by splitting each g at their mid-point. This results in a maximum of $|G_p| + 1$ candidate sets for the next iteration which happens when no pruning was done due to U^{acc} and if $|G_p|$ is odd.

Utility	Risk	k	Acceptable decrease in U	Best p	%Dec. in Risk
POI prediction	Home address inference	1	5	0.74	19.1
		5	5	0.62	9.2
		10	5	0.57	10.1
		1	10	0.81	23.1
		5	10	0.85	17.1
		10	10	0.79	18.1
	Re-identification threat	1	5	0.72	19.1
		5	5	0.53	16.1
		10	5	0.58	17.1
		1	10	0.92	28.4
		5	10	0.81	21.3
		10	10	0.83	20.1

TABLE A.1: Early stopping heuristic : POI@k

Utility	Risk	k	Acceptable decrease in U	Best p	%Dec. in Risk
Activity prediction	Home address inference	1	5	0.63	15.1
		5	5	0.52	12.2
		10	5	0.42	9.3
		1	10	0.79	21.2
		5	10	0.62	15.1
		10	10	0.56	13.5
	Re-identification threat	1	5	0.59	17.2
		5	5	0.52	13.1
		10	5	0.47	9.4
		1	10	0.69	21.8
		5	10	0.61	17.2
		10	10	0.52	13.7

TABLE A.2: Early stopping heuristic : Activity Prediction@k

A.3 Complexity Analysis

We envision the obfuscation to be performed offline by the data collector before sharing the location data with advertisers. Our obfuscation scheme requires computing features $\mathcal{F}(T_i)$ and inference of u_i and r_i for a trajectory T_i (or an obfuscation of it $\mathcal{P}(T_i)$) from a trained machine learning heuristic. Denote these inference times for a single consumer trajectory T_i as $O(F_i)$, $O(u_i)$ and $O(r_i)$. Note that these vary depending on the choice of privacy risk and utility function made by the data collector. To compute the estimates presented in Figure 4, we vary the grid parameter $p \in \{0, 0.1, \dots, 1\}$ and estimate $E(u_i)$ and $E(r_i)$ for twenty trials. This involves $20 \times 10 \times N \times O(F_i) \approx N \times O(F_i)$ time for feature computation. Once the features are built, these feed into the corresponding risk and utility estimation - $20 \times 10 \times N \times O(u_i) \times O(r_i) \approx N \times O(u_i) \times O(r_i)$. Hence the total time complexity is bounded by $O(N(F_i + u_i r_i))$.

1. In the case of Random Forests, which we employ for sensitive inference threat and in the sped-up heuristic for re-identification threat (Section 2), since $N \gg d$, the inference complexity is bounded by $O(N)$, hence $O(u_i) \approx 1$. For the POI prediction, the inference is again linear, to compute the nearest k locations for each consumer using a selection algorithm. Hence, the overall complexity is bounded by $O(NF_i)$ for both the privacy threats considered in the case of POI prediction.
2. In the activity prediction scenario, we employ a LSTM to quantify the

Utility	Risk	Clock Time (seconds)		
		Data	Full grid	Early stop
Next POI	Re-identification	100%	865	226
Next POI	Home address inference	100%	978	258
Activity prediction	Re-identification	100%	1390	312
Activity prediction	Home address inference	100%	1543	396
Next POI	Re-identification	50%	503	136
Next POI	Home address inference	50%	645	187
Activity prediction	Re-identification	50%	790	210
Activity prediction	Home address inference	50%	832	225

TABLE A.3: Clock time of the proposed heuristic

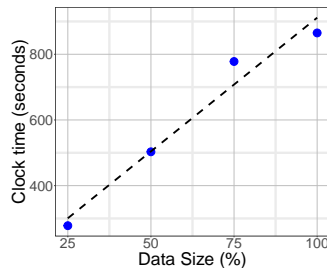


FIGURE A.1: Home address inference, POI prediction

utility. The inference time is linear in the number of points. While we do not need feature computation for the activity prediction, these still need to be computed to quantify consumer risk. Hence, the complexity is the same as earlier, bounded by $O(NF_i)$.

In the proposed early stopping heuristic, additional compute overhead arises from spanning across a finer grid of parameters, averaging over $M = 50$ repeated trials, until a stopping criterion is met. However, this overhead is offset since the estimates are computed on a subset of the data N_s . This is observed in Figure A.1 and Table A.3, where the early stopping heuristic is on an average four times faster than the fixed-point grid-based search.

We repeat all the experiments reported in the original paper with the proposed early stopping heuristic and report the resulting relative decreases in consumer risk and advertiser utility in Tables A.1, A.2.

A.4 Speed-up Heuristic.

While the re-identification risk can be exactly computed for a given $|\bar{T}_i|$, it is computationally inefficient with a complexity of $O(\left(\frac{T_i}{|\bar{T}_i}\right) \times N)$. To speed up the computation, we leverage a recent study (Pellungrini et al., 2018) that empirically shows the predictability of the re-identification risk for a given k using mobility features. The main idea is to learn a supervised algorithm, Random Forest, by building a set of mobility features similar to $\mathcal{F}(T)$ discussed in Section 1.4.1. We adopt this idea by further augmenting the mobility features with our consumer-consumer and consumer-location affinity features.

We then analytically compute the risks for a subset of the consumers and use the trained model to approximate the risks for the rest of the consumers (see Appendix A.6 for the technical details).

A.5 Utility Measurement

We compute the data utility under different obfuscations and by computing the performance of a neighborhood-based collaborative filtering recommendation model to accurately predict future consumer locations. To assess the accuracy of the predictions made, we treat the locations visited by each consumer in the fifth week as the ground truth and train the recommendation model to predict these locations.

Based on the consumer risks, we obfuscate T_{train} by varying $p \in \mathcal{P}$. We learn a neighborhood-based recommendation model (Bobadilla et al., 2011) by tuning the number of neighbors via five-fold cross-validation on the obfuscated training sample $\mathcal{P}(T_{train})$. The model is learned to rank the locations that a consumer is likely to visit during the fifth week of the observation period. That is, we build the features $\mathcal{F}(P(T_{train}))$ on first four weeks' data and tune the number of neighbors by using a grid of $\{5, 10, 25, 50, 100, 200\}$ to maximize the predictive accuracy. Then, we compute the data utility, $MAP@k$ and $MAR@k$, on T_{test} for $k = \{1, 5, 10\}$ to illustrate the efficacy of the proposed method. The learned recommendation model can be used to compute $MAP@k$ and $MAR@k$ for other values of k as well. Intuitively, $MAP@1$ and $MAR@1$, for example, represent an advertiser's utility to predict the next location most likely visited by a consumer in the fifth week based on the recommendation model learned on the obfuscated data. A key detail in the utility estimation is that we do not perform any obfuscation on T_{test} for any value of p , since our aim is to quantify the ability of obfuscated data, $\mathcal{P}(T_{train})$, to learn a consumer's true preference revealed in the unobfuscated test sample. Similar to the risk computation, we perform 20 trials for each p and report the mean and 95% confidence intervals of the utility metrics in Figure 1.4.

A.6 Model Choices in the Proposed Framework

We empirically justify our model choices in the proposed framework. All choices are made based by assessing the performance of different machine learning heuristics used in our framework on the unobfuscated data. First, in Figures A.2a and A.2b, we show the incremental benefit of the affinity features discussed in extracting the features $\mathcal{F}(T)$. Figure A.2a shows the accuracy of the Random Forest classifier in predicting each consumer's operating system. The model is regularized by performing a grid search on the maximum number of features $\{.25, .5, .75, 1\}$ and trees $\{50, 100, 200\}$ via five-fold cross-validation. The best performing model has an accuracy of 82% which indicates the success that a stalker would have in inferring the unpublished operating system of a consumer from the trajectory data. In Figure A.2b,

we plot the RMSE of the Random Forest regressor trained to predict home addresses.¹

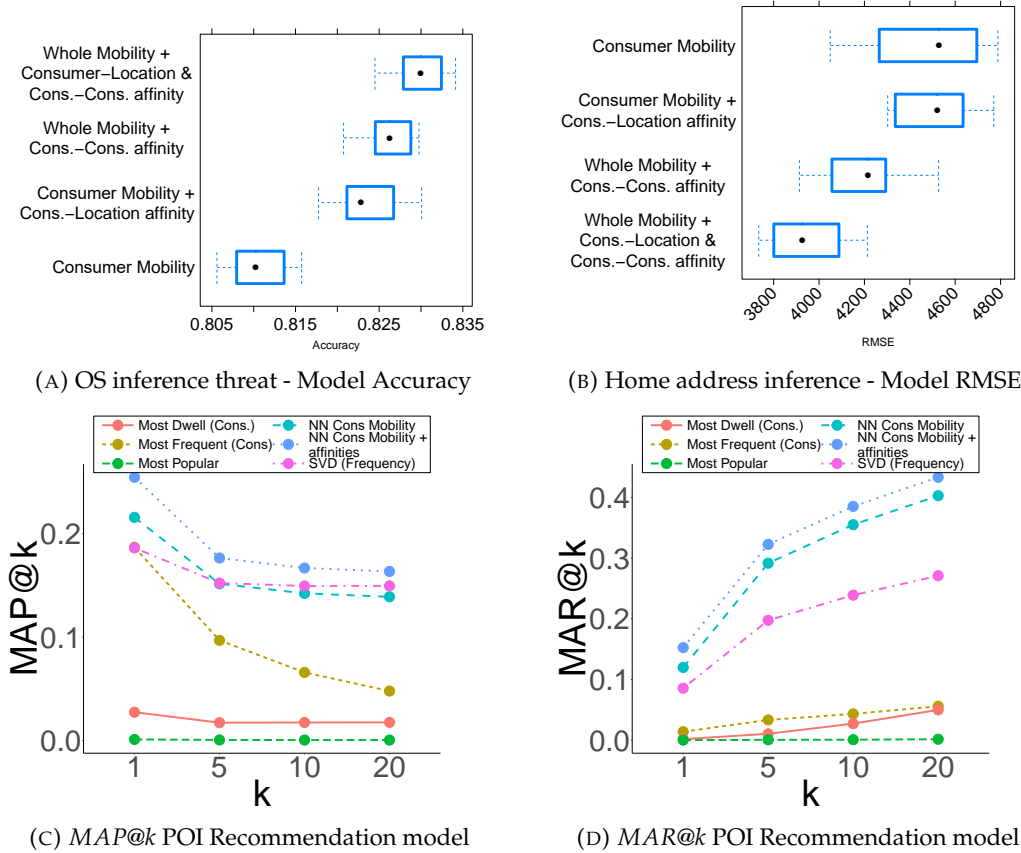


FIGURE A.2: Proposed framework model choices

Next, we learn two regression models to predict the Universal Transverse Mercator (UTM) transformed latitude and longitude of the home location with similar hyperparameter tuning as earlier. The error estimate is the Euclidean distance between the estimated and assigned home UTM coordinates. From the box plots of the re-sampled performance measures (Figures A.2a and A.2b), we notice that the consumer-consumer and consumer-location affinity features incrementally improve the performance of both proxy models learned. In Figures A.2c and A.2d, we visualize the $MAP@k$ and $MAR@k$ of the neighborhood-based recommendation model learned by tuning the number of neighbors.

We compare the performance with several baselines - recommendations based on the most popular locations (Most Popular), locations that the consumer spent the most time in (Most Dwell (consumer)), visited most frequently (Most Frequent (consumer)), and a singular value decomposition (SVD) on the consumer-location matrix populated with visit frequency. We observe that the NN based model performs better in both metrics compared to the baselines, justifying the choice. The RMSE, 3,900 meters \approx 2.46 miles indicates the success that a stalker would have in identifying a consumer's

¹We treat each consumer's most frequently visited location 10pm-6am as the ground truth of home location. The results remain robust across alternative time periods, e.g. 11pm-5am. We do not save these home locations to preserve consumer privacy.

home location from the unobfuscated data. Further, we also notice the incremental benefit of the affinity features in the recommendation performance (See NN consumer Mobility vs NN consumer Mobility + affinities in Figures A.2c and A.2d).

A.7 Additional Literature Review

Marketing research on consumer privacy falls into four main streams: consumer-, firm-, regulation-, and methodology- focused. We will concisely the first three streams here. The first stream takes on a consumers' perspective, and as a result, derives implications for firms to design privacy-friendly policies. For instance, a number of studies examine how consumers respond to privacy concerns or make privacy choices about privacy-intruding survey questions (Acquisti, John, and Loewenstein, 2012), platform provided privacy settings (Burtch, Ghose, and Wattal, 2015; Adjrid, Acquisti, and Loewenstein, 2018), online display ads that match website contents but with obtrusive format (Goldfarb and Tucker, 2011c; Goldfarb and Tucker, 2011b), or opt-in/out options of email marketing programs (Kumar, Zhang, and Luo, 2014). Other studies explore how normative and heuristic decision processes influence consumers' privacy decision making (Adjrid, Peer, and Acquisti, 2016). Overall, these studies point to positive effects of granting consumers enhanced controls over their own privacy, such as increasing their likelihood of responding to sensitive survey questions or click on personalized ads (Tucker, 2013). Interestingly, this stream of research also reveals that consumers behave in a way that reflects a "privacy paradox": claiming to care about their personal data yet more than willing to exchange the data for concrete benefits, such as convenience, personalization, or discounts (Acquisti and Grossklags, 2005; Chellappa and Sin, 2005; Awad and Krishnan, 2006; Xu et al., 2011; Ghose, 2017; Luo et al., 2014; Ghose, Li, and Liu, 2018), lower insurance premiums (Soleymanian, Weinberg, and Zhu, 2019), or a wider reach to audiences on social media for information acquisition or propagation (Adjrid, Acquisti, and Loewenstein, 2018). This paradox conversely indicates the potential for butler advertisers to leverage the newest mobile location data for geo-marketing to consumers in a mutually beneficial manner.

The second stream of literature assumes a firms' perspectives, often using a game-theoretic approach to reach normative implications of firms' privacy policies. For instance, Chellappa and Shivendu, 2010 derive an optimal design of personalization services for customers with heterogeneous privacy concerns. Gardete and Bart, 2018 propose an optimal choice of ad content and communication when the firm withholds the customers' private information. Conitzer, Taylor, and Wagman, 2012 reveal a monopoly's optimal cost of privacy for customers to remain anonymous. Hann et al., 2008 show that consumers' different actions toward preserving their privacy, such as address concealment or deflecting marketing, impact a firm's actions to either shifting marketing toward other consumers or reduce marketing overall. Adding competition to the picture, this stream of research also suggests optimal competitive strategies when profiting from disclosing customer information (Casadesus-Masanell and Hervas-Drane, 2015), or designing a B2B

market which preserves privacy to incentivize competitor participation (Kalvenes and Basu, 2006). Other studies have also conceptualized the differential importance of privacy to different platforms (Bart et al., 2005) and assessed the impact of data breaches on firms' financial performances (Martin, Borah, and Palmatier, 2017). Interestingly, this stream of research also demonstrates that firms, such as an ad network, do have innate incentives to preserve customer privacy even without privacy regulations (Rafieian and Yoganarasimhan, 2018).

The third stream of research focuses on privacy regulations. For example, these regulations are shown to impact firms' privacy-pertinent practices, technology innovations (Adjerid, Peer, and Acquisti, 2016) and adoptions (Miller and Tucker, 2009; Miller and Tucker, 2017), and consumers' responses to e.g. the do-no-call registry (Goh, Hui, and Png, 2015). European Union (EU)'s privacy policy is shown to reduce the effectiveness of online display ads (Goldfarb and Tucker, 2011a). Different components of a privacy law may also incur different effects, for instance, granting consumers controls over re-disclosure encourages genetic testing, whereas privacy notification deters it (Miller and Tucker, 2017).

A.8 Sensitive Attribute : Consumer Operating System

To further exemplify the robustness of our method, we repeat and report all the experiments discussed in Section 1.5 for another sensitive attribute - operating system of consumer's smartphone. Previous studies have shown a strong relationship between mobile operating system and consumer demographics in the context of mobile marketing (eMarketer, 2013).

In Figure A.3, we report the utility risk trade-off achieved by the proposed method. We note that the results remain qualitatively similar to our earlier findings of home address and re-identification threats. Quantitatively, in Figure A.3, we observe that a data collector can reduce the risk of inference by 10% without any decrease to advertiser's utility of POI prediction.

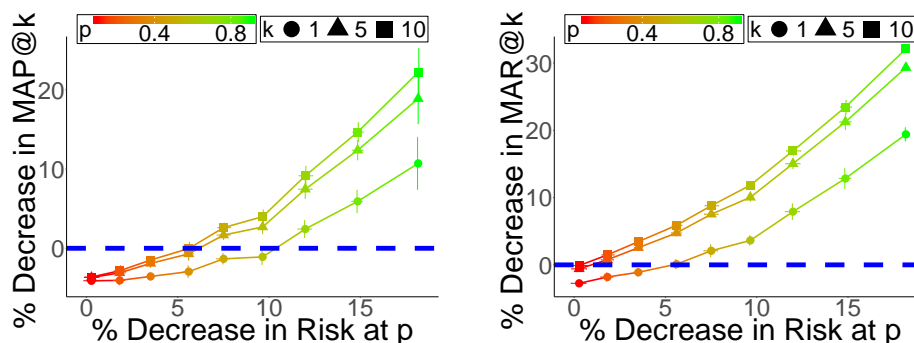


FIGURE A.3: Proposed framework - $MAP@k$ and $MAR@k$ for varying p , OS inference

In Figure A.4, we report the comparison of the proposed method to its ablations. The qualitative findings remain the same.

In Figure A.4, we compare the proposed method to rule based obfuscation schemes. We observe that the risk is reduced by $\approx 18\%$ (Figure A.3, $p = 0.9$, $k = 1$) compared to 25.49% when the timestamps are removed. However, this

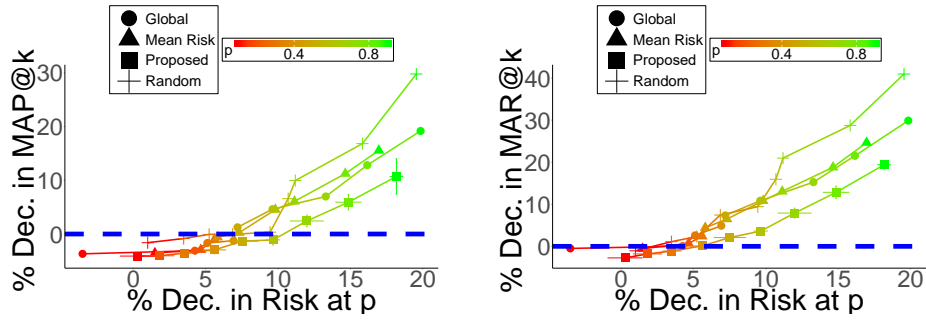


FIGURE A.4: Proposed framework vs risk-based obfuscations - $MAP@1$ and $MAR@1$, OS inference

Obfuscation rule	% Decrease Operating system risk	% Decrease Utility ($MAP@1$)	% Decrease Utility ($MAR@1$)
Remove Sleep hours	12.51	11.83	12.69
Remove Sleep and working hours	21.84	34.45	23.72
Remove time stamps	25.49	33.16	32.97

TABLE A.4: Alternative Schemes: Rule-based Obfuscation (Operating System Inference)

is achieved with a lesser decrease in the utility $\approx 10\%$ using the proposed framework when compared to 33%.

Obfuscation Method	% Decrease Operating system risk	% Decrease Utility (MAP@1)	% Decrease Utility (MAR@1)
GSUP ($P_{br} = 0.2$)	9.26	7.74	8.31
GSUP ($P_{br} = 0.5$)	3.11	4.49	3.42
LSUP ($P_{br} = 0.2$)	14.56	5.31	7.12
LSUP ($P_{br} = 0.5$)	4.01	-1.65	0.86

TABLE A.5: LSUP and GSUP comparison : OS inference (Green/Red indicate proposed framework provides a better/worse trade-off)

Finally, in Table A.5, we report the decrease in risk and utility for the recent suppression models and compare it to the trade-off provided by the proposed approach. We observe that in all of the four cases, the proposed method provides a better trade-off.

Appendix B

Social Determinants of Health

B.1 Location to Activity trajectories

Home and Work activity groups : Given T_i of an individual, we find the location where a consumer spends the most time from 1 AM - 5 AM on all days, map this location across T_i to activity *home*. A similar design to identify home location has been earlier used in the location data literature to address privacy concerns (Macha et al., 2019). To assign work, we exclude the *home* and locations within a 200m buffer around it (*not-home*) and proceed in the following sequence.

1. If the average time spent per day, across the observation period, at a *not-home* location is greater than 5 hours, indicating full-time work, we assign that location as consumer's *work*.
2. If the average time spent at a top *not-home* location is greater than 2 hours, indicating part-time work, we assign this location as *work* in T_i . If multiple locations satisfy this condition, indicating multiple part-time vocations, we assign all of these locations as *work*.
3. If a consumer spends less than 30 minutes at 3 or more *not-home* locations, indicating delivery behaviour, we assign such locations for a certain day as *work* to construct T_i .
4. Finally, if we are not able to identify a secondary *not-home* location where a consumer spent significant time in, we assume that the consumer does not have a steady job.

Other activity groups : To map the rest of the locations in T_i to activities, we identify a point of interest closest to the location using the Google Places API¹. The API returns a list of decorations (refer to second column of Table 3.1) which can help capture an individual's behavior (*gym, amusement_park, hair_care*), their consumption (*restaurant, meal_takeaway, cafe*) and their leisure activities (*art_gallery, spa, bowling_alley*) for each location. We aggregate decorations with similar semantics to construct thirteen more activities (first column in Table 3.1) in addition to home and work. To add a temporal context, we append each mapped activity with a coarser timestamp of t_j^i, c_j^i : 12 - 2 AM, 3 - 5 AM, 5 - 7 AM, 7 - 9 AM, 9 - 11 AM, 11 - 2 PM, 2 - 5 PM, 5 - 7 PM, 7 - 9 PM, 9 - 12 PM.

¹Google Places https://developers.google.com/places/web-service/supported_types

B.2 Author Topic Models Primer:

The Author Topic model (ATM), introduced by Rosen-Zvi et al., 2012 is a probabilistic generative model for documents that extends LDA Steyvers et al., 2004 to include authorship of documents. In ATM, each author is associated with a multinomial distribution over topics and each topic, like LDA, is associated with a multinomial distribution over words. By modeling the interests of authors, ATM enables us to establish what topics an author writes about, which authors are likely to have written documents similar to an observed document, and which authors produce similar work.

Figure 3.1 illustrates the generative process with a graphical model using plate notation. Shaded and unshaded circles indicate observed and latent variables respectively. An arrow indicates a conditional dependency between variables and plates (the boxes in Figure 3.1) indicate repeated sampling with the number of repetitions given by the variable in the bottom. In ATM, we observe both w and a_d , the set of authors of document d . When generating a document, an author is chosen at random ($Uniform(a_d)$) for each individual word in the document. The author picks a topic from their multinomial distribution over topics ($A \times K$ matrix, denoted by θ), and then samples a word from the multinomial distribution over words associated with that topic ($W \times K$ matrix, denoted by ϕ). This process is repeated for all words in the document until all the documents are created. Formally, the distributions of the unobserved variables are

$$\begin{aligned}
 P(\theta|\alpha) &\sim Dirichlet(\alpha) \\
 P(\phi|\beta) &\sim Dirichlet(\beta) \\
 P(z|x, \theta^{(x)}) &\sim Multinomial(\theta^{(x)}) \\
 P(w|z, \phi_{(z)}) &\sim Multinomial(\phi_{(z)}) \\
 P(x|a_d) &\sim Uniform(a_d)
 \end{aligned} \tag{B.1}$$

Note that the last equation simplifies to $x = a_d$; $|a_d| = 1$ in our lifestyle identification since each document would only comprise of a single consumer's day to day activities.

Gibbs Sampling and Estimation : The main objectives of ATM inference are to estimate the probability of generating w from topic k , $\phi_k^{(w)}$ and the probability of assigning topic k to a word generated by author a , $\theta_k^{(a)}$. More generally, for a given training corpus D_{train} , we need to estimate an approximation of the posterior distribution $P(\theta, \phi|z, x, D_{train}, \alpha, \beta)$, where $P(\theta, \phi|\alpha, \beta) = P(\theta|\alpha)P(\phi|\beta)$. Following the approach suggested in Rosen-Zvi et al., 2012, we first obtain an empirical sample based estimate of $P(z, x|D_{train}, \alpha, \beta)$ using Gibbs sampling for 1000 iterations (chaining). Next, we compute posterior estimates by leveraging the fact that Dirichlet and multinomial are conjugate distributions (Refer Eq 3.7).

B.3 D.C. Residents

Weekday Lifestyles: Figure B.1 visualizes the five identified weekday lifestyles and their corresponding activities for D.C. residents. Lifestyle 5 (denoted by

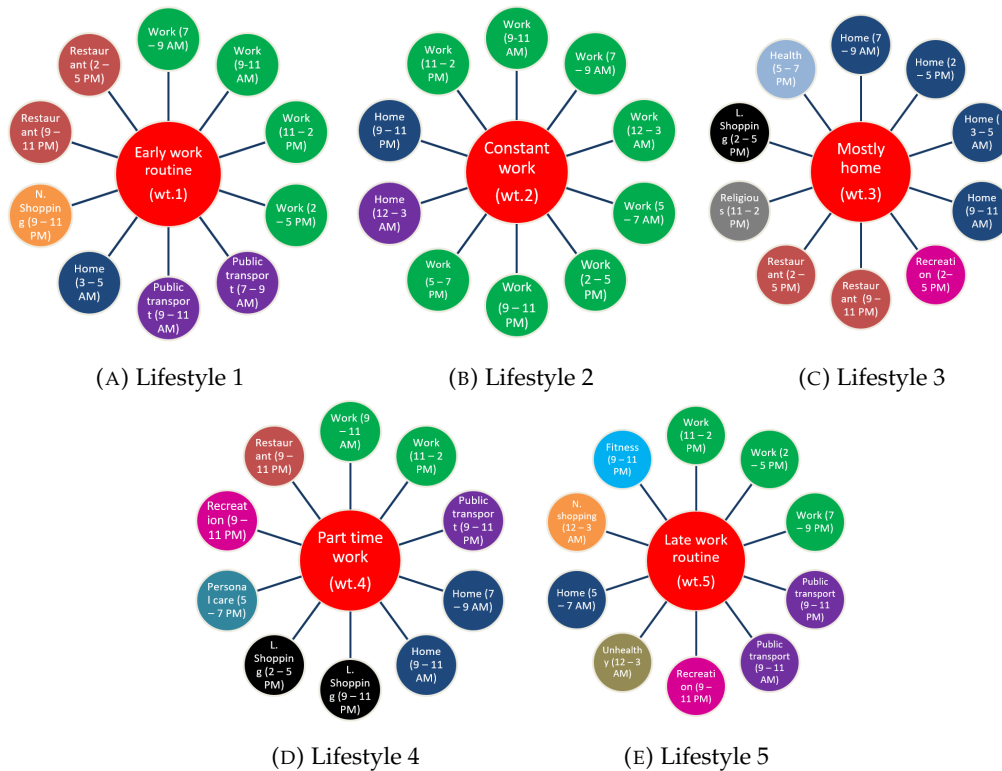


FIGURE B.1: DC Weekday Lifestyles

wt.5) characterizes a late work routine (*work* over 11 - 2 PM, 2 - 5 PM, 7 - 9 PM), commute via public transportation mornings and evenings (*publictransport* over 9 - 11 AM, 9 - 11 PM), late night recreation, unhealthy activities, fitness and necessity shopping (*recreation* in 9 - 11 PM, *fitness* 9 - 11 PM, *necessityshopping* 12 - 3 AM, *unhealthy.activities* 12 - 3 AM,). In contrast, lifestyle 1 (*wt.1*, with similar commute pattern, reveals an early work routine: *work* 7 - 9 AM, 9 - 11 AM, 11 - 2 PM, 2 - 5 PM, and consumption at restaurants during evenings and nights (*restaurant* 2 - 5 PM, 9 - 11 PM). Both lifestyles feature a steady full-time work routine, and work-fitness balance. Lifestyle *wt.4*, in comparison, indicates a part-time job (*work* 9 - 11 PM, 11 - 2 AM). Lifestyle 3 (*wt.3*) reveals a mostly at home routine while lifestyle 2 (*wt.2*) indicates multiple full-time/part-time jobs.

Weekend Lifestyles: Figure B.2 displays the top ten activities of the four weekend lifestyles for D.C. residents. Different from Baltimore, we observe that two lifestyles *wwt.1* (*work* 11 - 2 PM, 2 - 5 PM) and *wwt.4* (*work* 5 - 7 PM) with work activities, all other lifestyles suggest a non-work routine. Lifestyle *wwt.3* characterizes an early start weekend routine with fitness activities (*fitness* 9 - 11 AM) and *personal.care* afterwards (2 - 5 PM). In contrast, lifestyle *wwt.2* indicates consumption at restaurants later in the morning (*restaurant* 11 - 2 PM), with shopping and religious activities later in the evening (*leisure.shopping*, *necessity.shopping* 5 - 7 PM, *religious* 9 - 11 PM). Besides work on weekends, individuals in lifestyle *wwt.1* regularly consume at restaurants (*restaurant* 11 - 2 PM, 9 - 11 PM).

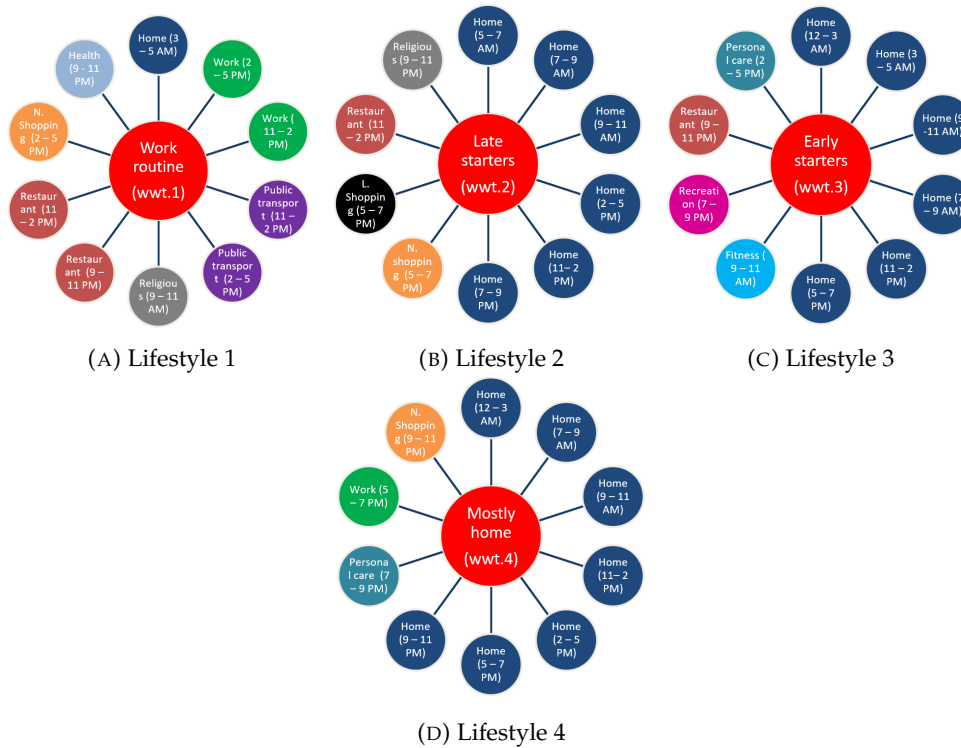


FIGURE B.2: DC Weekend Lifestyles

Model-free Evidence: Figure B.3 exhibits the histogram of the percentage of the 6,114 Baltimore residents with each lifestyle visiting medical facilities. Weekday lifestyles $wt.3$ and $wt.2$ have higher (4.14% and 2.92%, Figure B.3a) than average (2.58%) percentages of individuals visiting medical facilities. In contrast, lifestyle $wt.4$ has fewer than (1.94 %) average percentage of hospitalizations. Similarly, Figure B.3c reveals that weekend lifestyles $wwt.1$ and $wwt.2$ experience higher percentages of hospitalizations whereas lifestyle $wwt.4$ are sixty percent less likely.

In Figures B.3b, B.3d), we present the lift scores of weekend and weekday lifestyles. The top activities characterizing each lifestyle (Figures B.1, B.2) and their lift scores suggest that those who participate in *personal.care* activities on weekends ($wwt.4$) or weekdays ($wt.4$) are less likely (0.63 and 0.75, respectively) to have hospitalizations on average. On the other extreme, those with either busy, volatile work routines ($wt.2$) or no work routine ($wt.3$) on weekdays, are 1.12 and 1.60 times more likely to have hospitalizations (Figures B.3b, B.1; and people who either work ($wwt.1$) or are late starters ($wwt.2$) on weekends are 1.29 to 1.32 times more likely to have hospitalizations than average (Figures B.3d, B.2). Overall, the model-free evidence, similar to Baltimore residents, reveals heterogeneous rates of hospitalizations across different lifestyles.

Logit Analysis: Table B.1 (Column 4) indicate that those with $wwt.1$, $wt.3$, $wt.2$ have significantly higher odds of having a future hospitalization (1.31, 1.61, and 1.49 respectively) than average, after controlling for other social determinants. Similarly, lifestyles $wwt.4$ and $wt.4$ have significantly lower odds than average. These insights are qualitatively consistent with the model free

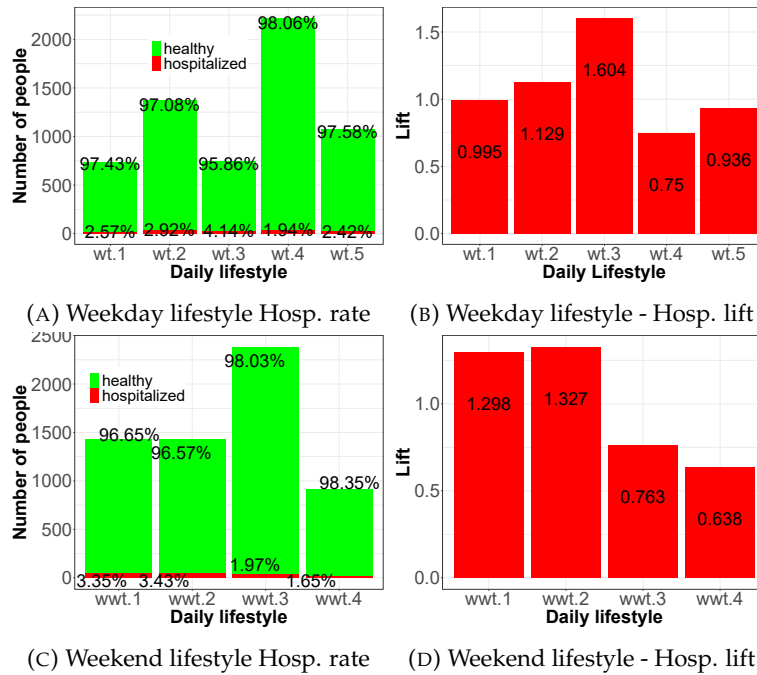


FIGURE B.3: (D.C.) Association with Hospitalization : Model free analysis (*hospitalization*)

evidence. Similar to Baltimore residents, we do not find any significant association between X_i^{access} , X_i^{demog} and future hospitalizations indicating that two individuals who live in the same neighborhood with similar social demographics, access to parks/fitness facilities, but with different lifestyles, will have different health risks. In Table B.2, we introduce total dwell time at healthy (fitness, personal care) and unhealthy activities into the regression and observe that *regularity* of personal care activities matters (lifestyle *wt.4*, *wwt.4*), instead of the total dwell time. In contrast to Baltimore, we do not observe that total time spent at unhealthy activities is significantly correlated to future hospitalization.

B.4 Robustness Checks

Sensitivity in Health Outcome : In Table 3.6 and 3.7, we reported the predictive performance of alternate definitions of health outcomes *hospitalization_night*, *hospitalization_alt*. To further showcase the robustness of our findings, we replicate our logit analysis for both Baltimore and D.C. residents (Tables B.3, B.4). We observe that both the qualitative and quantitative findings remain consistent with our key hospitalization variable *hospitalization*.

	<i>Dependent variable: hospitalization</i>				
	(1)	(2)	(3)	(4)	(5)
Weekend lifestyle 1 (wwt.1)	0.326** (0.150)	0.292* (0.151)	0.275* (0.159)	0.272* (0.164)	0.274* (0.165)
Weekend lifestyle 2 (wwt.2)	0.426** (0.175)	0.425** (0.176)	0.358* (0.181)	0.315 (0.184)	0.314 (0.185)
Weekend lifestyle 4 (wwt.4)	-0.473** (0.221)	-0.509** (0.221)	-0.417* (0.235)	-0.412* (0.235)	-0.410* (0.237)
Weekday lifestyle 2 (wt.2)	0.294 (0.193)	0.344* (0.195)	0.407* (0.198)	0.401* (0.206)	0.401* (0.210)
Weekday lifestyle 3 (wt.3)	0.387** (0.176)	0.387** (0.177)	0.414** (0.185)	0.481** (0.192)	0.483** (0.193)
Weekday lifestyle 4 (wt.4)	-0.359* (0.153)	-0.337* (0.155)	-0.350* (0.166)	-0.327* (0.170)	-0.327* (0.171)
Weekday lifestyle 5 (wt.5)	0.079 (0.188)	0.010 (0.191)	0.215 (0.203)	0.195 (0.207)	0.196 (0.218)
Accessibility metrics	✓	✓	✓	✓	✓
Mobility metrics	✓	✓	✓	✓	✓
Social Demographics	✓	✓	✓	✓	✓
Community Controls	✓	✓	✓	✓	✓
Observations	6,114	6,114	6,114	6,114	6,114
Log Likelihood	-484.149	-478.313	-453.994	-429.825	-398.986

Note: *p<0.1; **p<0.05; ***p<0.01

TABLE B.1: (D.C.) Hospitalization Logit Analysis

	<i>Dependent variable: hospitalization</i>		
	(1)	(2)	(3)
Weekend lifestyle 1 (wwt.1)	0.273* (0.167)	0.274* (0.167)	0.282* (0.172)
Weekend lifestyle 2 (wwt.2)	0.312 (0.186)	0.314 (0.187)	0.314 (0.193)
Weekend lifestyle 4 (wwt.4)	-0.411* (0.237)	-0.411* (0.242)	-0.411* (0.243)
Weekday lifestyle 2 (wt.2)	0.402* (0.211)	0.404* (0.211)	0.402* (0.213)
Weekday lifestyle 3 (wt.3)	0.484** (0.196)	0.492** (0.198)	0.496** (0.199)
Weekday lifestyle 4 (wt.4)	-0.329* (0.174)	-0.327* (0.178)	-0.328* (0.177)
Weekday lifestyle 5 (wt.5)	0.201 (0.217)	0.198 (0.218)	0.197 (0.218)
total_fitness_dwell	-0.002 (0.009)		
total_personalcare_dwell		-0.014 (0.014)	
total_unhealthyactivities_dwell			0.003 (0.004)
Other social determinants	✓	✓	✓
Observations	6,114	6,114	6,114
Log Likelihood	-378.824	-377.489	-378.868

Note: *p<0.1; **p<0.05; ***p<0.01

TABLE B.2: (D.C.) Hospitalization : Additional Logit Analysis

	Dependent variable:							
	hospitalization_alt				hospitalization_night			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Weekend lifestyle 1 (wwt.1)	0.220* (0.118)	0.216* (0.120)	0.197* (0.125)	0.170 (0.130)	0.213 (0.150)	0.205 (0.151)	0.101 (0.158)	0.149 (0.163)
Weekend lifestyle 2 (wwt.2)	-0.938** (0.190)	-0.920** (0.191)	-0.818** (0.197)	-0.784* (0.206)	-0.915** (0.263)	-0.894** (0.264)	-0.881** (0.270)	-0.798* (0.277)
Weekend lifestyle 4 (wwt.4)	-0.088 (0.116)	-0.093 (0.117)	-0.004 (0.124)	0.034 (0.129)	-0.018 (0.150)	-0.013 (0.150)	0.094 (0.158)	0.130 (0.163)
Weekday lifestyle 1 (wt.1)	-0.383*** (0.121)	-0.377*** (0.121)	-0.315** (0.128)	-0.288** (0.134)	-0.273* (0.151)	-0.282* (0.152)	-0.230 (0.159)	-0.191 (0.166)
Weekday lifestyle 2 (wt.2)	0.014 (0.129)	-0.010 (0.130)	-0.292** (0.136)	-0.287** (0.143)	0.093 (0.162)	0.086 (0.164)	-0.189 (0.170)	-0.108 (0.179)
Weekday lifestyle 4 (ww.4)	-0.377** (0.160)	-0.356** (0.162)	-0.311** (0.168)	-0.292* (0.177)	-0.516** (0.218)	-0.492** (0.219)	-0.442** (0.225)	-0.382* (0.240)
Weekday lifestyle 5 (wt.5)	0.714*** (0.114)	0.716*** (0.116)	0.627*** (0.122)	0.626*** (0.126)	0.966*** (0.136)	0.964*** (0.138)	0.886*** (0.144)	0.884*** (0.149)
Accessibility metrics	✓	✓	✓	✓	✓	✓	✓	✓
Mobility metrics	✓	✓	✓	✓	✓	✓	✓	✓
Social Demographics	✓	✓	✓	✓	✓	✓	✓	✓
Observations	4,528	4,528	4,528	4,528	4,528	4,528	4,528	4,528
Log Likelihood	-1,117.051	-1,108.370	-1,006.016	-923.317	-779.343	-775.067	-700.783	-643.706

Note:

*p<0.1; **p<0.05; ***p<0.01

TABLE B.3: (Baltimore) Logit Analysis : Robustness check for Hospitalization

	Dependent variable:							
	hospitalization_alt				hospitalization_night			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Weekend lifestyle 1 (wwt.1)	0.226** (0.102)	0.233** (0.103)	0.241** (0.107)	0.197* (0.111)	0.245* (0.142)	0.254* (0.143)	0.282* (0.148)	0.308** (0.155)
Weekend lifestyle 2 (wwt.2)	0.489** (0.124)	0.493** (0.125)	0.415** (0.129)	0.373** (0.132)	0.345** (0.168)	0.342** (0.168)	0.330* (0.173)	0.320* (0.179)
Weekend lifestyle 4 (wwt.4)	-0.429** (0.138)	-0.437** (0.138)	-0.420** (0.146)	-0.413** (0.150)	-0.456** (0.191)	-0.462** (0.192)	-0.384* (0.201)	-0.415* (0.210)
Weekday lifestyle 2 (wt.2)	-0.402** (0.135)	0.430** (0.137)	-0.458** (0.139)	-0.496** (0.143)	-0.435** (0.185)	-0.424** (0.187)	-0.467** (0.190)	-0.460** (0.197)
Weekday lifestyle 3 (wt.3)	0.279** (0.128)	0.267** (0.128)	0.332** (0.134)	0.381*** (0.138)	0.512*** (0.167)	0.508*** (0.167)	0.541*** (0.173)	0.582*** (0.181)
Weekday lifestyle 4 (wt.4)	-0.370** (0.105)	-0.373** (0.106)	-0.408*** (0.113)	-0.421*** (0.117)	-0.420*** (0.160)	-0.412** (0.162)	-0.422** (0.171)	-0.445** (0.181)
Weekday lifestyle 5 (wt.5)	0.187 (0.136)	0.188 (0.138)	0.196 (0.146)	0.207 (0.150)	0.205 (0.173)	0.215 (0.175)	0.225 (0.184)	0.211 (0.191)
Accessibility metrics	✓	✓	✓	✓	✓	✓	✓	✓
Mobility metrics	✓	✓	✓	✓	✓	✓	✓	✓
Social Demographics	✓	✓	✓	✓	✓	✓	✓	✓
Observations	6,114	6,114	6,114	6,114	6,114	6,114	6,114	6,114
Log Likelihood	-1,286.838	-1,283.772	-1,233.189	-1,160.998	-757.339	-755.076	-730.246	-674.633

Note:

*p<0.1; **p<0.05; ***p<0.01

TABLE B.4: (D.C.) Logit Analysis : Robustness check for Hospitalization

Bibliography

- Abul, Osman, Francesco Bonchi, and Mirco Nanni (2008). "Never walk alone: Uncertainty for anonymity in moving objects databases". In: *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*. Ieee, pp. 376–385.
- Acquisti, Alessandro and Jens Grossklags (2005). "Privacy and rationality in individual decision making". In: *IEEE security & privacy* 3.1, pp. 26–33.
- Acquisti, Alessandro, Leslie K John, and George Loewenstein (2012). "The impact of relative standards on the propensity to disclose". In: *Journal of Marketing Research* 49.2, pp. 160–174.
- Adjerid, Idris, Alessandro Acquisti, and George Loewenstein (2018). "Choice architecture, framing, and cascaded privacy choices". In: *Management Science*.
- Adjerid, Idris, Eyal Peer, and Alessandro Acquisti (2016). "Beyond the privacy paradox: Objective versus relative risk in privacy decision making". In: *Available at SSRN 2765097*.
- Agarwal, Vibhu et al. (2018). "Inferring physical function from wearable activity monitors: analysis of free-living activity data from patients with knee osteoarthritis". In: *JMIR mHealth and uHealth* 6.12, e11315.
- Aggarwal, Charu C (2005). "On k-anonymity and the curse of dimensionality". In: *Proceedings of the 31st international conference on Very large data bases*. VLDB Endowment, pp. 901–909.
- Aggarwal, Charu C. (2013). *Outlier Analysis*. Springer. ISBN: 978-1-4614-6396-2.
- Aggarwal, Charu C. et al. (1999). "Fast Algorithms for Projected Clustering." In: *SIGMOD*, pp. 61–72.
- Aggarwal, Gagan et al. (2005). "Approximation algorithms for k-anonymity". In: *Journal of Privacy Technology (JOPT)*.
- Agrawal, Rakesh et al. (1998). "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications." In: *SIGMOD*, pp. 94–105.
- Andrews, Michelle et al. (2016). "Mobile ad effectiveness: Hyper-contextual targeting with crowdedness". In: *Marketing Science* 35.2, pp. 218–233.
- Angiulli, Fabrizio, Fabio Fassetti, and Luigi Palopoli (Apr. 21, 2009). "Detecting outlying properties of exceptional objects." In: *ACM Trans. Database Syst.* 34.1. URL: <http://dblp.uni-trier.de/db/journals/tods/tods34.html#AngiulliFP09>.
- (2013). "Discovering Characterizations of the Behavior of Anomalous Subpopulations." In: *IEEE TKDE* 25.6, pp. 1280–1292.
- Apple (2012). *Apple Has Quietly Started Tracking iPhone Users Again, And It's Tricky To Opt Out, 2012*. <http://www.businessinsider.com/ifa-apples-iphone-tracking-in-ios-6-2012-10>.
- (2014). *Requesting Permission*. <https://developer.apple.com/design/human-interface-guidelines/ios/app-architecture/requesting-permission/>.
- (2016). *iOS 10 to Feature Stronger "Limit Ad Tracking" Control, 2016*. <https://pfp.org/2016/08/02/ios-10-feature-stronger-limit-ad-tracking/>.
- Ashbrook, Daniel and Thad Starner (2003). "Using GPS to learn significant locations and predict movement across multiple users". In: *Personal and Ubiquitous computing* 7.5, pp. 275–286.

- Awad, Naveen Farag and Mayuram S Krishnan (2006). "The personalization privacy paradox: an empirical evaluation of information transparency and the willingness to be profiled online for personalization". In: *MIS quarterly*, pp. 13–28.
- Bao, Ling and Stephen S Intille (2004). "Activity recognition from user-annotated acceleration data". In: *International conference on pervasive computing*. Springer.
- Bart, Yakov et al. (2005). "Are the drivers and role of online trust the same for all web sites and consumers? A large-scale exploratory empirical study". In: *Journal of marketing* 69.4, pp. 133–152.
- Bayardo, Roberto J and Rakesh Agrawal (2005). "Data privacy through optimal k-anonymization". In: *null*. IEEE, pp. 217–228.
- Ben-Zeev, Dror et al. (2014). "Feasibility, acceptability, and preliminary efficacy of a smartphone intervention for schizophrenia". In: *Schizophrenia bulletin* 40.6, pp. 1244–1253.
- Berger, Paul D and Nada I Nasr (1998). "Customer lifetime value: Marketing models and applications". In: *Journal of interactive marketing* 12.1, pp. 17–30.
- Bobadilla, Jesus et al. (2011). "A framework for collaborative filtering recommender systems". In: *Expert Systems with Applications* 38.12, pp. 14609–14623.
- Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.
- Bruner, Gordon C and Anand Kumar (2007). "Attitude toward location-based advertising". In: *Journal of interactive advertising* 7.2, pp. 3–15.
- Buchbinder, Niv et al. (2014). "Submodular Maximization with Cardinality Constraints." In: *SODA*, pp. 1433–1452.
- Burch, Gordon, Anindya Ghose, and Sunil Wattal (2015). "The hidden cost of accommodating crowdfunder privacy preferences: a randomized field experiment". In: *Management Science* 61.5, pp. 949–962.
- Casadesus-Masanell, Ramon and Andres Hervas-Drane (2015). "Competing with privacy". In: *Management Science* 61.1, pp. 229–246.
- Chawla, Nitesh V et al. (2002). "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16, pp. 321–357.
- Chellappa, Ramnath K and Shivendu Shivendu (2010). "Mechanism design for "free" but "no free disposal" services: The economics of personalization under privacy concerns". In: *Management Science* 56.10, pp. 1766–1780.
- Chellappa, Ramnath K and Raymond G Sin (2005). "Personalization versus privacy: An empirical examination of the online consumer's dilemma". In: *Information technology and management* 6.2-3, pp. 181–202.
- Chen, Min, Xuan Tan, and Rema Padman (2020). "Social determinants of health in electronic health records and their impact on analysis and risk prediction: A systematic review". In: *JAMIA* 27.11, pp. 1764–1773.
- Chen, Rui, Gergely Acs, and Claude Castelluccia (2012). "Differentially private sequential data publication via variable-length n-grams". In: *Proceedings of the 2012 ACM conference on Computer and communications security*. ACM, pp. 638–649.
- Chen, Rui et al. (2013). "Privacy-preserving trajectory data publishing by local suppression". In: *Information Sciences* 231, pp. 83–97.
- Chen, Tianqi and Carlos Guestrin (2016). "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, pp. 785–794.
- Cheng, Chun Hung, Ada Wai-Chee Fu, and Yi Zhang (1999). "Entropy-based Subspace Clustering for Mining Numerical Data." In: *KDD*, pp. 84–93.
- Chiuve, Stephanie E et al. (2012). "Alternative dietary indices both strongly predict risk of chronic disease". In: *The Journal of nutrition* 142.6, pp. 1009–1018.

- Chow, Chi-Yin and Mohamed F Mokbel (2011). "Trajectory privacy in location-based services and data publication". In: *ACM Sigkdd Explorations Newsletter* 13.1, pp. 19–29.
- Clark, Peter and Tim Niblett (1989). "The CN2 induction algorithm". In: *Machine learning* 3.4, pp. 261–283.
- Clifton, Chris and Tamir Tassa (2013). "On syntactic anonymity and differential privacy". In: *2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*. IEEE, pp. 88–93.
- Cockerham, William C, Thomas Abel, and Günther Lüschen (1993). "Max Weber, formal rationality, and health lifestyles". In: *Sociological Quarterly* 34.3, pp. 413–425.
- Cohen, William W (1995). "Fast effective rule induction". In: *Machine Learning Proceedings 1995*. Elsevier, pp. 115–123.
- Conitzer, Vincent, Curtis R Taylor, and Liad Wagman (2012). "Hide and seek: Costly consumer privacy in a market with repeat purchases". In: *Marketing Science* 31.2, pp. 277–292.
- Dang, Xuan Hong et al. (2013). "Local Outlier Detection with Interpretation." In: *ECML/PKDD*, pp. 304–320.
- Dang, Xuan Hong et al. (2014). "Discriminative features for identifying and interpreting outliers." In: *ICDE*, pp. 88–99.
- Dave, Vacha, Saikat Guha, and Yin Zhang (2012). "Measuring and fingerprinting click-spam in ad networks." In: *SIGCOMM*. ACM, pp. 175–186.
- Dawadi, Prafulla N, Diane J Cook, and Maureen Schmitter-Edgecombe (2016). "Modeling patterns of activities using activity curves". In: *Pervasive and mobile computing* 28, pp. 51–68.
- De Jong, Martijn G, Rik Pieters, and Jean-Paul Fox (2010). "Reducing social desirability bias through item randomized response: An application to measure underreported desires". In: *Journal of Marketing Research* 47.1, pp. 14–27.
- De Lathauwer, Lieven, Bart De Moor, and Joos Vandewalle (2000). "A multilinear singular value decomposition". In: *SIAM journal on Matrix Analysis and Applications* 21.4, pp. 1253–1278.
- Deng, Houtao (2014). "Interpreting tree ensembles with intrees". In: *arXiv preprint arXiv:1408.5456*.
- Dubé, Jean-Pierre et al. (2017). "Competitive price targeting with smartphone coupons". In: *Marketing Science* 36.6, pp. 944–975.
- Dwork, Cynthia and Jing Lei (2009). "Differential privacy and robust statistics". In: *Proceedings of the forty-first annual ACM symposium on Theory of computing*. ACM, pp. 371–380.
- Eagle, Nathan and Alex Sandy Pentland (2009). "Eigenbehaviors: Identifying structure in routine". In: *Behavioral Ecology and Sociobiology* 63.7, pp. 1057–1066.
- Eagle, Nathan, Alex Sandy Pentland, and David Lazer (2009). "Inferring friendship network structure by using mobile phone data". In: *Proceedings of the national academy of sciences* 106.36, pp. 15274–15278.
- eMarketer (2013). *US Smartphone OS Race Still Close, as Men, Younger Users Favor Android*. <https://www.emarketer.com/article.aspx?R=1009961&RewroteTitle=1>.
- Fanaee-T, Hadi and João Gama (2015). "Eigenevent: An algorithm for event detection from complex data streams in syndromic surveillance". In: *Intelligent Data Analysis* 19.3, pp. 597–616.
- Fang, Zheng et al. (2015). "Contemporaneous and delayed sales impact of location-based mobile promotions". In: *Information Systems Research* 26.3, pp. 552–564.

- Farrahi, Katayoun and Daniel Gatica-Perez (2011). "Discovering routines from large-scale human locations using probabilistic topic models". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 2.1, pp. 1–27.
- Fisher, Anne G and Kristin Bray Jones (2012). *Assessment of Motor and Process Skills Volume II-User Manual*. Three Star Press.
- Fong, Nathan M, Zheng Fang, and Xueming Luo (2015a). "Geo-conquesting: Competitive locational targeting of mobile promotions". In: *Journal of Marketing Research* 52.5, pp. 726–735.
- (2015b). "Real-Time Mobile Geo-Conquesting Promotions". In: *Journal of Marketing Research*.
- Fong, Ruth C and Andrea Vedaldi (2017). "Interpretable explanations of black boxes by meaningful perturbation". In: *arXiv preprint arXiv:1704.03296*.
- Freeman, WJ, AJ Weiss, and KC Heslin (2016). "Overview of US Hospital Stays in 2016: Variation by Geographic Region". In: *Rockville, MD: Agency for Healthcare Research and Quality*.
- Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2001). *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York.
- Friedman, Jerome H, Bogdan E Popescu, et al. (2008). "Predictive learning via rule ensembles". In: *The Annals of Applied Statistics* 2.3, pp. 916–954.
- Fung, Benjamin C. M. et al. (June 2010). "Privacy-preserving Data Publishing: A Survey of Recent Developments". In: *ACM Comput. Surv.* 42.4, 14:1–14:53. ISSN: 0360-0300. DOI: [10.1145/1749603.1749605](https://doi.org/10.1145/1749603.1749605). URL: <http://doi.acm.org/10.1145/1749603.1749605>.
- Gamberger, Dragan and Nada Lavrac (2002). "Expert-guided subgroup discovery: Methodology and application". In: *Journal of Artificial Intelligence Research* 17, pp. 501–527.
- Gao, Sheng et al. (2014). "Balancing trajectory privacy and data utility using a personalized anonymization model". In: *Journal of Network and Computer Applications* 38, pp. 125–134.
- García-Olmos, Luis et al. (2019). "Development of a predictive model of hospitalization in primary care patients with heart failure". In: *Plos one* 14.8, e0221434.
- Gardete, Pedro M and Yakov Bart (2018). "Tailored cheap talk: The effects of privacy policy on ad content and market outcomes". In: *Marketing Science* 37.5, pp. 733–752.
- Garfinkel, Robert, Ram Gopal, and Paulo Goes (2002). "Privacy protection of binary confidential data against deterministic, stochastic, and insider threat". In: *Management Science* 48.6, pp. 749–764.
- Gharan, Shayan Oveis and Jan Vondrak (2011). "Submodular Maximization by Simulated Annealing." In: *SODA*. SIAM, pp. 1098–1116.
- Ghose, Anindya (2017). *TAP: Unlocking the mobile economy*. MIT Press.
- Ghose, Anindya, Beibei Li, and Siyuan Liu (2018). "Mobile targeting using customer trajectory patterns". In: *Management Science* Forthcoming.
- Ghosh, Shalini et al. (2016). "Contextual lstm (clstm) models for large scale nlp tasks". In: *arXiv preprint arXiv:1602.06291*.
- Goh, Khim-Yong, Kai-Lung Hui, and Ivan PL Png (2015). "Privacy and marketing externalities: Evidence from do not call". In: *Management Science* 61.12, pp. 2982–3000.
- Goldfarb, Avi and Catherine Tucker (2011a). "Online display advertising: Targeting and obtrusiveness". In: *Marketing Science* 30.3, pp. 389–404.
- (2011b). "Rejoinder—Implications of "Online Display Advertising: Targeting and Obtrusiveness"". In: *Marketing Science* 30.3, pp. 413–415.

- Goldfarb, Avi and Catherine E Tucker (2011c). "Privacy regulation and online advertising". In: *Management science* 57.1, pp. 57–71.
- Gonzalez, Marta C, Cesar A Hidalgo, and Albert-Laszlo Barabasi (2008). "Understanding individual human mobility patterns". In: *nature* 453.7196, p. 779.
- Görnitz, Nico, Marius Kloft, and Ulf Brefeld (2009). "Active and Semi-supervised Data Domain Description." In: *ECML/PKDD*, pp. 407–422.
- Greer, Danielle M et al. (2014). "Milwaukee Health Report 2013: Health disparities in Milwaukee by socioeconomic status". In:
- Guenther, Patricia M, Jill Reedy, and Susan M Krebs-Smith (2008). "Development of the healthy eating index-2005". In: *Journal of the American Dietetic Association* 108.11, pp. 1896–1901.
- Guo, Mingming et al. (2015). "In-network trajectory privacy preservation". In: *ACM Computing Surveys (CSUR)* 48.2, pp. 1–29.
- Hann, Il-Horn et al. (2008). "Consumer privacy and marketing avoidance: A static model". In: *Management Science* 54.6, pp. 1094–1103.
- Hara, Satoshi and Kohei Hayashi (2016). "Making tree ensembles interpretable". In: *arXiv preprint arXiv:1606.05390*.
- Hastie, Trevor et al. (2009). "Multi-class adaboost". In: *Statistics and its Interface* 2.3, pp. 349–360.
- Hayes, Tamara L et al. (2008). "Unobtrusive assessment of activity patterns associated with mild cognitive impairment". In: *Alzheimer's & Dementia* 4.6, pp. 395–405.
- He, Jingrui and Jaime G. Carbonell (2010). "Co-selection of Features and Instances for Unsupervised Rare Category Analysis." In: *SDM*, pp. 525–536.
- He, Jingrui, Hanghang Tong, and Jaime G. Carbonell (2010). "Rare Category Characterization." In: *ICDM*, pp. 226–235.
- He, Xi et al. (2015). "DPT: differentially private trajectory synthesis using hierarchical reference systems". In: *Proceedings of the VLDB Endowment* 8.11, pp. 1154–1165.
- Herrera, Franciso et al. (2011). "An overview on subgroup discovery: foundations and applications". In: *Knowledge and information systems* 29.3, pp. 495–525.
- Hilton, C Beau et al. (2020). "Personalized predictions of patient outcomes during and after hospitalization using artificial intelligence". In: *NPJ Digital Medicine* 3.1, pp. 1–8.
- Hoffman, Donna L, Thomas P Novak, and Marcos Peralta (1999). "Building consumer trust online". In: *Communications of the ACM* 42.4, pp. 80–85.
- Hothorn, Torsten, Kurt Hornik, and Achim Zeileis (2015). "ctree: Conditional inference trees". In: *The Comprehensive R Archive Network*, pp. 1–34.
- Hu, Tianran et al. (2016). "Mining shopping patterns for divergent urban regions by incorporating mobility data". In: *Proceedings of the 25th ACM CIKM*, pp. 569–578.
- Huo, Zheng et al. (2012). "You can walk alone: trajectory privacy-preserving through significant stays protection". In: *International conference on database systems for advanced applications*. Springer, pp. 351–366.
- Huynh, Tâm, Mario Fritz, and Bernt Schiele (2008). "Discovery of activity patterns using topic models". In: *Proceedings of the 10th international conference on Ubiquitous computing*, pp. 10–19.
- Hwang, Ren-Hung, Yu-Ling Hsueh, and Hao-Wei Chung (2013). "A novel time-obfuscated algorithm for trajectory privacy protection". In: *IEEE Transactions on Services Computing* 7.2, pp. 126–139.
- ICSI (2004). "Going beyond clinical walls : Solving Complex Problem". In:

- Joumard, Isabelle et al. (2010). "Health status determinants: lifestyle, environment, health care resources and efficiency". In: *Environment, Health Care Resources and Efficiency (May 27, 2010)*. OECD Economics Department Working Paper 627.
- Kalvenes, Joakim and Amit Basu (2006). "Design of robust business-to-business electronic marketplaces with guaranteed privacy". In: *Management Science* 52.11, pp. 1721–1736.
- Keller, Fabian, Emmanuel Müller, and Klemens Böhm (2012). "HiCS: High Contrast Subspaces for Density-Based Outlier Ranking." In: *ICDE*, pp. 1037–1048.
- Keller, Fabian et al. (2013). "Flexible and adaptive subspace search for outlier analysis." In: *CIKM*. ACM, pp. 1381–1390.
- Kelsey (2018). *US Local Mobile Local Social Ad Forecast*. <https://shop.biakelsey.com/product/2018-u-s-local-mobile-local-social-ad-forecast/>.
- Klösgen, Willi (1996). "Explora: A multipattern and multistrategy discovery assistant". In: *Advances in knowledge discovery and data mining*. American Association for Artificial Intelligence, pp. 249–271.
- Klösgen, Willi and Michael May (2002). "Census data mining—an application". In: *Proceedings of the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), Helsinki, Finland*.
- Knorr, Edwin M. and Raymond T. Ng (1999). "Finding Intensional Knowledge of Distance-Based Outliers." In: *VLDB*, pp. 211–222.
- Koh, Pang Wei and Percy Liang (2017). "Understanding black-box predictions via influence functions". In: *arXiv preprint arXiv:1703.04730*.
- Kopp, Martin, Tomáš Pevný, and Martin Holena (2014). "Interpreting and clustering outliers with sapling random forests". In: *ITAT*.
- Kriegel, Hans-Peter, Peer Kröger, and Arthur Zimek (2009). "Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering". In: *ACM Trans. Knowl. Discov. Data* 3.1, pp. 1–58. ISSN: 1556-4681. DOI: <http://doi.acm.org/10.1145/1497577.1497578>.
- Kriegel, Hans-Peter et al. (2005). "A Generic Framework for Efficient Subspace Clustering of High-Dimensional Data." In: *ICDM*.
- Kriegel, Hans-Peter et al. (2009). "Outlier Detection in Axis-Parallel Subspaces of High Dimensional Data." In: *PAKDD*, pp. 831–838.
- Kriegel, Hans-Peter et al. (2012). "Outlier Detection in Arbitrarily Oriented Subspaces." In: *ICDM*, pp. 379–388.
- Kumar, V, Xi Zhang, and Anita Luo (2014). "Modeling customer opt-in and opt-out in a permission-based marketing context". In: *Journal of Marketing Research* 51.4, pp. 403–419.
- Kuo, Chia-Tung and Ian Davidson (2016). "A Framework for Outlier Description Using Constraint Programming." In: *AAAI*, pp. 1237–1243.
- Lakkaraju, Himabindu et al. (2017). "Interpretable and Explorable Approximations of Black Box Models." In: *CoRR* abs/1707.01154.
- Lazarevic, Aleksandar and Vipin Kumar (2005). "Feature bagging for outlier detection." In: *KDD*, pp. 157–166.
- Lee, Kyumin, Brian David Eoff, and James Caverlee (2011). "Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter." In: *ICWSM*.
- Li, Chen, Houtan Shirani-Mehr, and Xiaochun Yang (2007). "Protecting individual information against inference attacks in data publishing". In: *International Conference on Database Systems for Advanced Applications*. Springer, pp. 422–433.
- Li, Chenxi et al. (2017). In: *Marketing Science* 36.5, pp. 762–779.

- Li, Tiancheng et al. (2012). "Slicing: A new approach for privacy preserving data publishing". In: *IEEE transactions on knowledge and data engineering* 24.3, pp. 561–574.
- Li, Xiao-Bai and Sumit Sarkar (2009). "Against classification attacks: A decision tree pruning approach to privacy protection in data mining". In: *Operations Research* 57.6, pp. 1496–1509.
- Li, Yanping et al. (2018). "Impact of healthy lifestyle factors on life expectancies in the US population". In: *Circulation* 138.4, pp. 345–355.
- Liaw, Andy, Matthew Wiener, et al. (2002). "Classification and regression by randomForest". In: *R news* 2.3, pp. 18–22.
- Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou (2008). "Isolation Forest." In: *ICDM*.
- Loekito, Elsa and James Bailey (2008). "Mining influential attributes that capture class and group contrast behaviour." In: *CIKM*. ACM, pp. 971–980. URL: <http://dblp.uni-trier.de/db/conf/cikm/cikm2008.html#LoekitoB08>.
- Loewenstein, George, Troyen Brennan, and Kevin G Volpp (2007). "Asymmetric paternalism to improve health behaviors". In: *Jama* 298.20, pp. 2415–2417.
- Logan, Beth et al. (2007). "A long-term evaluation of sensing modalities for activity recognition". In: *International conference on Ubiquitous computing*. Springer, pp. 483–500.
- Luo, Xueming et al. (2014). "Mobile targeting". In: *Management Science* 60.7, pp. 1738–1756.
- Macha, Meghanath et al. (2019). "Perils of Location Tracking? Personalized and Interpretable Privacy Preservation in Consumer Mobile Trajectories". In: *Working Paper, Carnegie Mellon University*.
- Machanavajhala, Ashwin, Johannes Gehrke, and Michaela Götz (2009). "Data publishing against realistic adversaries". In: *Proceedings of the VLDB Endowment* 2.1, pp. 790–801.
- Machanavajhala, Ashwin et al. (2006). " ℓ -Diversity: Privacy Beyond κ -Anonymity". In: *null*. IEEE, p. 24.
- Martin, Kelly D, Abhishek Borah, and Robert W Palmatier (2017). "Data privacy: Effects on customer and firm performance". In: *Journal of Marketing* 81.1, pp. 36–58.
- Mayer, Charles S and Charles H White Jr (1969). "The law of privacy and marketing research". In: *Journal of Marketing* 33.2, pp. 1–4.
- Meyerson, Adam and Ryan Williams (2004). "On the complexity of optimal k-anonymity". In: *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, pp. 223–228.
- Micenková, Barbora et al. (2013). "Explaining Outliers by Subspace Separability." In: *ICDM*, pp. 518–527.
- Miller, Amalia R and Catherine Tucker (2009). "Privacy protection and technology diffusion: The case of electronic medical records". In: *Management Science* 55.7, pp. 1077–1093.
- (2017). "Privacy protection, personalized medicine, and genetic testing". In: *Management Science* 64.10, pp. 4648–4668.
- Miller, Richard D and Ted Frech (2002). "The productivity of health care and pharmaceuticals: quality of life, cause". In:
- Moise, Gabriela, Jörg Sander, and Martin Ester (2006). "P3C: A Robust Projected Clustering Algorithm." In: *ICDM*, pp. 414–425.
- Molitor, Dominik et al. (2019). "Measuring the effectiveness of location-based advertising: A randomized field experiment". In: *Available at SSRN* 2645281.

- Montavon, Grégoire, Wojciech Samek, and Klaus-Robert Müller (2017). "Methods for interpreting and understanding deep neural networks". In: *Digital Signal Processing*.
- Morse, Steven, Marta C Gonzalez, and Natasha Markuzon (2016). "Persistent cascades: Measuring fundamental communication structure in social networks". In: *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, pp. 969–975.
- Mukherjee, Arjun et al. (2013). "What Yelp Fake Review Filter Might Be Doing?" In: *ICWSM*.
- Müller, Emmanuel, Matthias Schiffer, and Thomas Seidl (2011). "Statistical selection of relevant subspace projections for outlier ranking." In: *ICDE*, pp. 434–445.
- Müller, Emmanuel et al. (2008). "OutRank: ranking outliers in high dimensional data." In: *ICDE Workshops*, pp. 600–603.
- Müller, Emmanuel et al. (2009). "Relevant Subspace Clustering: Mining the Most Interesting Non-redundant Concepts in High Dimensional Data." In: *ICDM*. IEEE, pp. 377–386.
- Müller, Emmanuel et al. (2012). "Outlier Ranking via Subspace Analysis in Multiple Views of the Data." In: *ICDM*, pp. 529–538.
- Muralidhar, Krishnamurthy and Rathindra Sarathy (2006). "Data shuffling—A new masking approach for numerical data". In: *Management Science* 52.5, pp. 658–670.
- Nixon, John and Philippe Ulmann (2006). "The relationship between health care expenditure and health outcomes". In: *The European Journal of Health Economics* 7.1, pp. 7–18.
- Paavilainen, Paula et al. (2005). "Circadian activity rhythm in demented and non-demented nursing-home residents measured by telemetric actigraphy". In: *Journal of sleep research* 14.1, pp. 61–68.
- Pappalardo, Luca, Salvatore Rinzivillo, and Filippo Simini (2016). "Human mobility modelling: exploration and preferential return meet the gravity model". In: *Procedia Computer Science* 83, pp. 934–939.
- Parsons, L., E. Haque, and H. Liu (2004). "Subspace clustering for high dimensional data: a review". In: 6.1, pp. 90–105.
- Pelekis, Nikos et al. (2011). "Privacy-aware querying over sensitive trajectory data". In: *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 895–904.
- Pelleg, Dan and Andrew Moore (2000). "X-means: Extending K-means with Efficient Estimation of the Number of Clusters". In: *ICML*, pp. 727–734.
- Pellungrini, Roberto et al. (2018). "A data mining approach to assess privacy risk in human mobility data". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 9.3, p. 31.
- Pevný, Tomáš and Martin Kopp (2014). "Explaining anomalies with sapling random forests". In: *ITAT*.
- Pew (2016). *More Americans using smartphones for getting directions, streaming TV*. <http://www.pewresearch.org/fact-tank/2016/01/29/us-smartphone-use/>.
- (2018). *Americans' complicated feelings about social media in an era of privacy concerns*. <https://www.pewresearch.org/fact-tank/2018/03/27/americans-complicated-feelings-about-social-media-in-an-era-of-privacy-concerns/>.
- Qian, Yi and Hui Xie (2015). "Drive more effective data-based innovations: enhancing the utility of secure databases". In: *Management Science* 61.3, pp. 520–541.
- Rafieian, Omid and Hema Yoganarasimhan (2018). "Targeting and privacy in mobile advertising". In:

- Ramachandram, Dhanesh and Graham W Taylor (2017). "Deep multimodal learning: A survey on recent advances and trends". In: *IEEE Signal Processing Magazine* 34.6, pp. 96–108.
- Regulation, General Data Protection Regulation (2016). "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46". In: *Official Journal of the European Union (OJ)* 59.1-88, p. 294.
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). "Why should i trust you?: Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 1135–1144.
- Riboni, Daniele and Claudio Bettini (2012). "Private context-aware recommendation of points of interest: An initial investigation". In: *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*. IEEE, pp. 584–589.
- Rissanen, J. (1978). "Modeling By Shortest Data Description". In: *Automatica* 14, pp. 465–471.
- Robben, Saskia, Margriet Pol, and Ben Kröse (2014). "Longitudinal ambient sensor monitoring for functional health assessments: a case study". In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pp. 1209–1216.
- Robben, Saskia et al. (2012). "How is grandma doing? Predicting functional health status from binary ambient sensor data". In: *2012 AAAI Fall Symposium Series*.
- Rosen-Zvi, Michal et al. (2012). "The author-topic model for authors and documents". In: *arXiv preprint arXiv:1207.4169*.
- Rosenberg, Barry L et al. (2016). "Quantifying geographic variation in health care outcomes in the United States before and after risk-adjustment". In: *PLoS One* 11.12.
- Saeb, Sohrab et al. (2015). "Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study". In: *Journal of medical Internet research*.
- Samarati, Pierangela (2001). "Protecting respondents identities in microdata release". In: *IEEE transactions on Knowledge and Data Engineering* 13.6, pp. 1010–1027.
- Samarati, Pierangela and Latanya Sweeney (1998). "Generalizing data to provide anonymity when disclosing information". In: *PODS*. Vol. 98. Citeseer, p. 188.
- Sandıkçı, Burhaneddin et al. (2013). "Alleviating the patient's price of privacy through a partially observable waiting list". In: *Management Science* 59.8, pp. 1836–1854.
- Schneider, Matthew J et al. (2018). "A Flexible Method for Protecting Marketing Data: An Application to Point-of-Sale Data". In: *Marketing Science* 37.1, pp. 153–171.
- Sequeira, Karlton and Mohammed Javeed Zaki (2004). "SCHISM: A New Approach for Interesting Subspace Mining." In: *ICDM*, pp. 186–193.
- Sievert, Carson and Kenneth Shirley (2014). "LDavis: A method for visualizing and interpreting topics". In: *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pp. 63–70.
- Silverman, Bernard W (2018). *Density estimation for statistics and data analysis*. Routledge.
- Soleymanian, Miremad, Charles B Weinberg, and Ting Zhu (2019). "Sensor Data and Behavioral Tracking: Does Usage-Based Auto Insurance Benefit Drivers?" In: *Marketing Science*.

- Statista (2018). *Share of smartphone users that use an Apple iPhone in the United States from 2014 to 2019*. <https://www.statista.com/statistics/236550/percentage-of-us-population-that-own-a-iphone-smartphone/>.
- Steyvers, Mark et al. (2004). "Probabilistic author-topic models for information discovery". In: *Proceedings of the 10th ACM SIGKDD*.
- Sun, Feng-Tso et al. (2014). "Nonparametric discovery of human routines from sensor data". In: *2014 IEEE international conference on pervasive computing and communications (PerCom)*. IEEE, pp. 11–19.
- Tax, David M. J. and Robert P. W. Duin (Dec. 8, 2005). "Support Vector Data Description." In: *Machine Learning* 54.1, pp. 45–66.
- Taylor, Kyle and Laura Silver (2019). *Smartphone Ownership Is Growing Rapidly Around the World, but Not Always Equally*. <http://www.pewglobal.org/2019/02/05/smartphone-ownership-is-growing-rapidly-around-the-world-but-not-always-equally/>.
- Terrovitis, Manolis, Nikos Mamoulis, and Panos Kalnis (2008). "Privacy-preserving anonymization of set-valued data". In: *Proceedings of the VLDB Endowment* 1.1, pp. 115–125.
- Terrovitis, Manolis et al. (2017). "Local suppression and splitting techniques for privacy preserving publication of trajectories". In: *IEEE Trans. Knowl. Data Eng* 29.7, pp. 1466–1479.
- Thompson, Stuart A. and Charlie Warzel (2019). *Twelve Million Phones, One Dataset, Zero Privacy*. <https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html>.
- Tucker, Catherine E (2013). "Social networks, personalized advertising, and privacy controls". In: *Journal of Marketing Research* 50.5, pp. 546–562.
- Valentino-Devries, Jennifer et al. (2018). *Your Apps Know Where You Were Last Night, and They're Not Keeping It Secret*. <https://www.nytimes.com/interactive/2018/12/10/business/location-data-privacy-apps.html?smid=renytimes>.
- Verge (2019). *Android Q leak reveals system-wide dark mode and bigger emphasis on privacy*. <https://www.theverge.com/2019/1/16/18185763/android-q-leak-dark-mode-new-privacy-settings>.
- Vreeken, Jilles, Matthijs Van Leeuwen, and Arno Siebes (2011). "Krimp: mining itemsets that compress". In: *Data Mining and Knowledge Discovery* 23.1, pp. 169–214.
- Wachs, Theodore D et al. (2015). "Issues in the timing of integrated early interventions: contributions from nutrition, neuroscience, and psychological research." In:
- Wang, Dashun et al. (2011). "Human mobility, social ties, and link prediction". In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. Acm, pp. 1100–1108.
- Wang, Ke, Benjamin CM Fung, and S Yu Philip (2007). "Handicapping attacker's confidence: an alternative to k-anonymization". In: *Knowledge and Information Systems* 11.3, pp. 345–368.
- Wedel, Michel and PK Kannan (2016). "Marketing analytics for data-rich environments". In: *Journal of Marketing* 80.6, pp. 97–121.
- Wernke, Marius et al. (2014). "A classification of location privacy attacks and approaches". In: *Personal and ubiquitous computing* 18.1, pp. 163–175.
- Williams, Nathalie E et al. (2015). "Measures of human mobility using mobile phone records enhanced with GIS data". In: *PloS one* 10.7, e0133630.

- Wrobel, Stefan (1997). "An algorithm for multi-relational discovery of subgroups". In: *European Symposium on Principles of Data Mining and Knowledge Discovery*. Springer, pp. 78–87.
- Xu, David Jingjun (2006). "The influence of personalization in affecting consumer attitudes toward mobile advertising in China". In: *Journal of Computer Information Systems* 47.2, pp. 9–19.
- Xu, Heng et al. (2011). "The personalization privacy paradox: An exploratory study of decision making process for location-aware marketing". In: *Decision support systems* 51.1, pp. 42–52.
- Xue, Andy Yuan et al. (2013). "Destination prediction by sub-trajectory synthesis and privacy protection against such prediction". In: *2013 IEEE 29th international conference on data engineering (ICDE)*. IEEE, pp. 254–265.
- Yang, Dingqi, Bingqing Qu, and Philippe Cudre-Mauroux (2018). "Privacy-Preserving Social Media Data Publishing for Personalized Ranking-Based Recommendation". In: *IEEE Transactions on Knowledge and Data Engineering*.
- Yarovoy, Roman et al. (2009). "Anonymizing moving objects: How to hide a mob in a crowd?" In: *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*. ACM, pp. 72–83.
- Yuan, Quan, Gao Cong, and Aixin Sun (2014). "Graph-based point-of-interest recommendation with geographical and temporal influences". In: *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pp. 659–668.
- Yuan, Quan et al. (2013). "Time-aware point-of-interest recommendation". In: *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pp. 363–372.
- Zhang, Haopeng, Yanlei Diao, and Alexandra Meliou (2017). "EXstream: Explaining Anomalies in Event Stream Monitoring." In: *EDBT*, pp. 156–167.
- Zheng, Jiangchuan and Lionel M Ni (2012). "An unsupervised framework for sensing individual and cluster behavior patterns from human mobile data". In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pp. 153–162.
- Zheng, Yu, Xing Xie, and Wei-Ying Ma (2010). "Geolife: A collaborative social networking service among user, location and trajectory." In: *IEEE Data Eng. Bull.* 33.2, pp. 32–39.