From Private Location Data to Public Good

by

Alex Berke

B.A., Brown University (2013)

Submitted to the Program in Media Arts and Sciences
in partial fulfillment of the requirements for the degree of

Master of Science in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2020

Author .....
Program in Media Arts and Sciences
August 7, 2020

Certified by .....
Kent Larson
Principal Research Scientist
Thesis Supervisor

Accepted by .....
Tod Machover
Academic Head, Program in Media Arts and Sciences

From Private Location Data to Public Good

by

Alex Berke

Submitted to the Program in Media Arts and Sciences
on August 7, 2020, in partial fulfillment of the
requirements for the degree of
Master of Science in Media Arts and Sciences

Abstract

This thesis was written in the midst of the COVID-19 pandemic as location datasets became crucial sources of information to address the global health emergency. The subject of this thesis is how location data collected from mobile devices can be used to benefit the public and preserve individuals' privacy. The work presented in this thesis directly addresses the public health emergency as well as how these datasets can serve the public beyond the time of crisis. For example this thesis explores privacy-preserving technologies that use data collected from personal devices to scale contact tracing efforts. This is in order to stymie disease transmission as well as stem the adoption of privacy-violating technologies that were initially deployed by governments contending with COVID-19. The work in this thesis also leverages a high-precision and up-to-date location dataset collected from millions of smartphones across the U.S. to better understand the impacts of COVID-19 on communities and human behaviors. This includes developing new metrics to improve the monitoring and modeling of disease transmission. This thesis also explores strategies using machine learning models to generate privacy-preserving synthetic location data that can retain the utility of real location data and supplement traditional survey datasets.

Surveys collected by government agencies and research institutions often produce datasets and knowledge that serve as public goods. This thesis frames the ongoing collection of location data as an ongoing population survey. The ethics of data collection are beyond the scope of this work. Instead this thesis shows how location data which primarily benefits private industry can also benefit the public from whom it is sourced, in ways similar to traditional survey data, and protect individuals' privacy.

Thesis Supervisor: Kent Larson
Title: Principal Research Scientist

From Private Location Data to Public Good

by
Alex Berke

This thesis has been reviewed and approved by the following committee members

Kent Larson .....
Thesis Advisor
Principal Research Scientist, MIT Media Lab, City Science

Esteban Moro .....
Thesis Reader
Visiting Professor, MIT Media Lab

Ronan Doorley .....
Thesis Reader
Research Contractor, City Science, MIT Media Lab
Adjunct Assistant Professor, Trinity College Dublin

Acknowledgments

This work represents a time when I stepped away from what I had already learned how to do and took on new challenges. Writing this thesis was an exercise in reflecting on what I have learned, and my gratitude for opportunities to do so.

Thank you to my advisor Kent Larson for providing me intellectual and creative space to learn. Even when we did not discuss technical details of my explorations, he had a knack for asking the right high-level questions to direct my focus. His intellectual curiosity and ambition are a model for my own; I look forward to his guidance as we continue exploring together. Thank you to Ronan Doorley and Esteban Moro. Together we took on academic endeavors that were new to me - I learned from their expertise and this thesis would not have come to be what it is without them. I am fortunate to call them collaborators and look forward to continuing to learn from them. This includes learning from Ronan's superior ability to listen and think deeply about new problems and approach them with academic rigor. And from Esteban's inexhaustible energy and excitement in the quest of discovery and in what big data can help us understand. Thank you to collaborators Michiel Bakker, Matt Groh, Bernardo Bulle, and Dan Calacci. I learned from each of them and can only hope all of my future collaborators are as morally grounded, curious, and fun. Thank you to Lena Abdallah. Our collaboration and lessons learned together led to work presented in chapter 3 of this thesis. Thank you to Jason Nawyn and Thomas Sanchez. Even projects that are pursued out of passion can come to resemble work. Our collaborations are not featured in this thesis, but these inspired people provided me necessary reminders, in times of stress, of the passion that brought us to our work. And thank you to everyone else in the City Science group who I have had the pleasure of collaborating with, or simply sharing insights or space with (physical in safer times, virtual otherwise). Thank you to my (nontechnical) family who have offered unwavering support through each of my intellectual and professional endeavors, despite often not understanding them. Their love created a home from which I could safely wander.

Contents

Chapter 1

Introduction

There is an ongoing and ubiquitous collection of high-precision location data from mobile devices. As these devices continue to become increasingly pervasive parts of modern life, the datasets collected from them will become increasingly powerful tools. Location datasets are passively collected through a variety of mobile applications and they serve a flourishing data economy. They are highly valuable to the companies who collect them and are highly useful to those who pay to use them. Common use cases include improved advertisement targeting, company analytics, and other means for private profit.

Yet the true value and utility of location data has not yet been fully realized, as location datasets also have the potential to better serve the public.

Government agencies and research organizations commonly collect surveys and use survey data to better understand and provide benefit to the populations from whom the data are sourced[1]. Often the resulting datasets are released as public goods. The ongoing collection of data from mobile devices can be considered an ongoing population survey. And this new data source can better serve many of the same use cases as traditional survey data, as well as provide for new ones.

[1]For example, the U.S. Census Bureau oversees the collection of a variety of surveys and makes the data publicly available. The datasets range from information about housing supply, businesses, income and employment levels, as well as population demographics and social characteristics. This information informs planning decisions such as where to locate job training centers, build roads and schools, or provide services for the elderly [1]. The second part of this thesis includes examples of how publicly available survey data about trips between geographic regions inform disease transmission models [2].

This became urgently clear in the midst of the COVID-19 pandemic as location datasets became crucial sources of information to address the global health crisis. Many researchers used outdated and otherwise limited survey data to understand and model the response to the crisis. Location data collected from mobile devices provided a higher-resolution and more real-time source of information for these same purposes as well as for contact tracing. These use cases are the focus of the second part of the thesis.

There is precedent for publicly releasing mobility data collected from the public, such as through the use of transportation systems[2]. Datasets can serve as powerful tools and democratizing data in this way expands the possible uses and benefits they can serve [7]. The location datasets that are the subject of this work can serve as even more powerful tools than these other sources because they are higher precision and their continuous collection makes them more complete.

However, many of the same qualities that make this new data source highly useful also present privacy risks. Location data can reveal information about where people live, work, and frequent, and other places they went and when they went there. These data can represent people's activities and behaviors. This information is both useful and sensitive. For example, sensitive information such as religious affiliation, sexual preferences, medical history or political activity could be exposed [8]. Common strategies to anonymize location datasets do not effectively remove privacy risks [9]. Other methods used to curtail the privacy risks of these data involve aggregation or otherwise reducing the amount of information within the data, as doing so reduces the likelihood that an individual within the dataset could be re-identified, or reduces the likelihood that sensitive information about them could be recovered. However, methods that reduce the level of information in the data can also reduce the utility of the data. There is often then a trade-off between the utility and privacy of such datasets. The following chapters of this thesis further detail and address these issues.

This thesis is about using location data while addressing privacy risks. In particular, it is about using location data collected via mobile devices, which primarily serve private interests, in order to benefit the public from whom it is sourced.

[2]Many local governments publicly release mobility data collected through use of public infrastructure. A few common examples include data about bike-share trips, taxi rides, and metro station entrances and exits [3, 4, 5, 6]. This is often due to explicit open data policies made by cities and transportation agencies.

The contributions of this thesis are presented in two parts: The first part of the work is about generating new privacy-preserving *synthetic* location datasets. The second part is about using *real* location data to address the global COVID-19 health crisis, with a focus on privacy.

In the first part of the work, I present a system designed to generate synthetic data that represent real location data. By generating realistic data, the utility of the real data can be retained, while privacy risks for real users can be circumvented because the generated data represent the location histories of a synthetic population. Previous works have also generated synthetic datasets; I cover these related works as well as their limitations. In short, they do not cover the full scope of the work I present. The models developed in this work take information about a population distribution as input to generate realistic location data for that given population. A location based services dataset is used to train the system's models and evaluate the generated output by how well it preserves the utility of the real data, as well as preserves privacy. Through this work I encountered a shortcoming in the privacy literature. I address this by developing privacy criteria, which builds upon related computational privacy works, to evaluate privacy for synthetic mobility data.

In this work I approach the problem of generating realistic location data by exploiting the patterns inherent in individuals' mobility data, in order to generate data that retain these patterns. Patterns in mobility data reflect the routines of everyday life. The sequences of where people go, and when they go there, are laden with spatial and temporal relationships. By recognizing these patterns we can model an individual's mobility data as analogous to words in a sentence, or notes in music, and approach the data generation problem with machine learning models that have been successful in text and music generation.

In the second part of the thesis I demonstrate important ways location data can benefit the public, namely by serving as a tool to address the COVID-19 global health crisis. This part also discusses the resulting privacy implications and ways to mitigate them.

From the beginning of the crisis, contact tracing emerged as a useful strategy to limit disease transmission. However, many of the early successful contact tracing outcomes were due to using location data in ways that jeopardized privacy and freedom. In chapter 5 I demonstrate how more privacy-preserving technologies can be built and still deliver useful information

with the immediacy required by the crisis. Chapter 4 provides a more comprehensive analysis of the various ways contact tracing technologies can use location data, and the trade-offs with respect to privacy and effectiveness.

Chapters 6 and 7 are about using location data to better study and monitor the response to the COVID-19 pandemic. My collaborators and I leverage a high-resolution location dataset to detect when people come into contact with each other. The aggregate "contacts" metrics we derive can protect privacy while still supplying the precise information we use to measure "social distancing" behaviors.

Through these contributions this thesis demonstrates the value location data can serve as a public good while respecting privacy.

Part I

GENERATING NEW MOBILITY DATASETS

Chapter 2

Motivation

Official statistics are traditionally collected through government surveys that ask respondents questions. The resulting datasets are often publicly published so that organizations and researchers beyond the agencies that collected the information can use them. The datasets, as well as research insights they lead to, serve as public goods.

Common examples of such surveys in the United States are the decennial census, which counts the entire population, and the American Community Survey, which targets a sample of the population to provide population estimates more frequently [10, 11]. Governments also collect travel surveys, which are central to the motivation of this work.

These datasets are considered authoritative sources, but they are severely limited when compared to new sources of information. Surveys are costly [12, 13] and collected infrequently, and the data often suffer from small samples with reporting bias [14, 15]. Surveys rely on respondents answering a set of preconceived questions, limiting the amount of information they can provide, and they rely on the accuracy of respondents' answers (i.e. they are subject to recall bias [16, 17]). They also can only report on a snapshot in time.

In contrast, location data are continuously collected from personal mobile devices. This process creates up-to-date datasets representing large samples of the population, where the accuracy of the data are not dependent on the memory of the "surveyed" population. These datasets are generated by multiple sources. For example, location data can come from

devices probing for Wi-Fi access points, or from posts by social media app users, such as with geotagged tweets or check-ins at restaurants. The datasets motivating this work are from mobile network operators and location based services. These datasets are already amassed and accessible by single entities and their collection process is ubiquitous and passive. Location data are collected by mobile network operators when phones connect with cell towers. Location based services (LBS) data are collected from smartphones by software in a variety of applications, including when applications are running in the background. While data from Wi-Fi or social apps may be limited to a smaller number of places, such as near Wi-Fi access points, or places where social app users chose to post, data can be collected by mobile network operators and location based services anywhere a user might have network coverage. And unlike data from sources like social apps, these data are collected passively, without device users taking explicit actions.

Datasets from mobile phones have been identified by governments and researchers as inexpensive tools to produce or update population estimates, or when used in combination with traditional surveys, reduce sample sizes, reduce costs, and reduce the respondent burden [18, 19]. Countries such as Estonia and Indonesia are already using data from mobile phones as part of the regular production of official statistics [20].

Our particular focus is on mobility data traditionally collected from travel surveys. Mobility data from these surveys inform transit engineers, decision makers, and researchers about the travel behaviors of the population. Mobile phone data can provide clear benefits over the data collected by these surveys, and has already been used by departments of transportation in several states of the U.S.[1] The National Household Travel Survey (NHTS) provides an explicit example for our motivation.

The NHTS is collected by the U.S. federal government[2] and is "the authoritative source on the travel behavior of the American public" [23]. Yet the information it provides is limited to only one randomly chosen day of travel and it is collected only once about every 8 years. Furthermore, the tiny sample size and high nonresponse bias of target survey respondents

[1] Use cases for mobile phone data by U.S. DOTs have ranged from measuring and signaling fluctuations of traffic speed [21], inferring traffic on major highways, and creating what are called origin-destination (O-D) matrices for the purposes of transportation modeling [22].

[2] The agency that oversees the NHTS is the Federal Highway Administration and the data collection and reporting is contracted out to a third party firm. This has not always been the case.

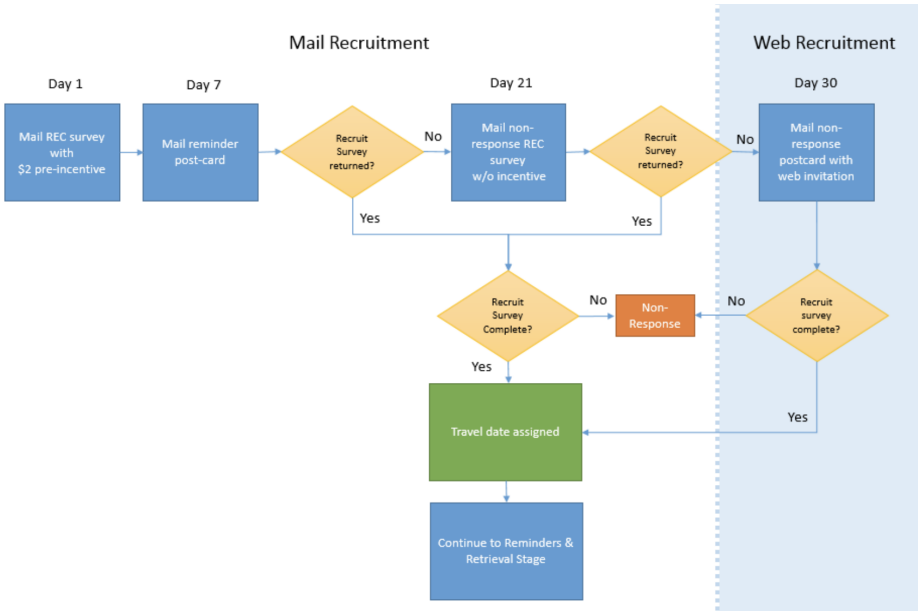


Figure 2-1: Flowchart summarizing the recruitment process for 2017 NHTS survey respondents, from figure 2-1 of the 2017 NHTS Data User Guide [24].

negatively affect the accuracy of the analyses that use this data. For example, the most recent survey (2017) includes only 804 people from the Boston Core Based Statistical Area, an area with a population estimate of 4,840,000 (i.e. approximately 0.017% of the population). So few survey respondents cannot accurately represent the overall population.

The NHTS reaches target respondents by sending survey packets to their mailing addresses (7.5% of packets are reported dropped), and recipients must then go through a multipart process to complete the survey, for which they receive a minor cash incentive[3] [24]. (See figure 2-1 for a diagram summarizing the recruitment process for 2017 survey respondents.) Unlike previous survey years, only a small fraction of target respondents (15.6%) completed the surveys[4] (see table 2.1). Coupling this statistic with the bothersome survey process and

[3]The NHTS incentive plan incrementally rewards participation. In the 2017 survey design, each sampled address received a $2 cash incentive in the initial recruitment mailing. The travel log package sent to each recruited household contained a $5 cash incentive. Finally, when the entire household completed the retrieval survey the household received an additional $20.

[4]The methods to collect survey data have changed over the years. The earliest versions of the NHTS, conducted in 1969, 1977, and 1983, were administered face-to-face using Census Bureau staff. These earlier surveys were also all conducted as retrospectives (e.g. a recall of the household's travel 'yesterday') rather than the recent method that tells respondents ahead of time the day to report on. To improve coverage and keep costs within reason, the 1990 survey was conducted by telephone, using a Random-Digit Dialing sample frame and Computer-Aided Telephone Interviewing. By 2009, however, there was real concern about the representativeness of the sampled population, which only included land-line telephone numbers. The 2017 NHTS used an address-based sample frame and a two-stage collection process.

survey year	1977	1983	1990	1995	2001	2009	2017
response rate	85.3%	94%	73%	37%	41%	19.8%	15.6%

Table 2.1: Travel survey response rates over time. Adapted from table 3-1 of the 2017 NHTS Data User Guide [24].

limited incentives makes clear that the resulting data suffer from nonresponse bias[5].

Despite these limitations, NHTS datasets are used as primary sources of information for research involving traffic safety, congestion, the environment, energy consumption, demographic trends, bicycle and pedestrian studies, and transit planning [26].

To overcome small sample sizes, researchers typically combine NHTS data with other sources that contain aggregate population estimates, such as census data, using statistical methods such as iterative proportional fitting[6]. This procedure generates a synthetic population that is more representative of the true population. However, it is still based on the original NHTS sample and limited by its potential bias and lack of diverse information.

Synthetic populations generated from NHTS data are also used to simulate the mobility behaviors of synthetic agents in agent-based models (ABMs). In particular, this is often done by researchers in the Media Lab's City Science research group, which is where this thesis work takes place.

Our lab also obtained access to a location based services (LBS) dataset that was provided by a location intelligence firm[7]. The dataset is representative of LBS datasets collected by similar and competitor companies. Data are collected daily and continuously without relying on respondents' memory or attention. This is in contrast to the most recent NHTS that asked respondents to report on their travel histories for just one day in 2017. The LBS data also represent a larger population sample than the NHTS. For example, our LBS dataset reports data for approximately 2.7% of the Boston area population, versus the 0.017% from the NHTS. (See the appendix section A.1 about data representativeness.)

[5]The organization that collects the survey acknowledges the issue of nonresponse bias [25].

[6]The iterative proportional fitting procedure is more commonly referred to as "raking" in survey statistics, such as with the weighting recommendations released with the NHTS [25].

[7]The LBS data was provided through a special program where the company grants access to anonymous, privacy-compliant location-based data for academic research and humanitarian initiatives related to human mobility.

However the LBS data also has limitations. While it represents a significantly larger sample of the population than the NHTS, it still does not represent the entire population. In addition, data collected from mobile devices and applications suffer from their own sources of bias, in that they only represent people using the devices or applications. Researchers often use other sources, such as official statistics from surveys, to better understand the degree of this bias. And similarly to NHTS data, census data are used to handle the small sample sizes. Later portions of this thesis work present examples of using these data sources together in such ways. Data from traditional surveys are also used to validate research results that are produced with mobile phone data [22, 27, 28].

For these reasons, data from traditional surveys play important roles in the use of data collected from mobile phones and these data sources can play complementary roles. Survey data can better inform the use of mobile phone data while data from mobile phones can add valuable information to supplement survey data.

Yet there are additional and important limitations for data collected from personal devices, such as the LBS data obtained by City Science: these datasets are highly sensitive. Public use of these datasets would present privacy risks for the device users from whom the data were collected and could detract value from the firms that collect them.

This part of the thesis approaches these issues with the generation of privacy-preserving synthetic data. In particular, by designing models, that are trained on real LBS data, to generate data that realistically represent the real data while sufficiently varying from it. In addition, the models can use population information from the official census in order to generate location data representing the full population. For example, for each person the ACS reports living in each census tract, the models can generate corresponding data.

The synthetic mobility datasets that are produced can be used to simulate the behaviors of the synthetic agents in agent-based models. They can also support the other kinds of research traditionally supported by travel surveys. By retaining the aggregate statistical properties of the real data, they can provide insights and answers for researchers' queries.

When the synthetic data sufficiently vary from the real data, the privacy of real users can be preserved. At the same time, the companies that are in the business of collecting the

data can retain value. By preserving their exclusive access to the real, and possibly higher precision data, they can continue to sell information for commercial uses that only the real data can provide (e.g. whether customers who were presented with an advertisement were more likely to then enter a store[8]).

Synthetic mobility datasets that are based on real LBS data but sufficiently vary from it, might then be made publicly available[9]. They can then complement travel surveys. And like traditional surveys, they can serve as public goods.

[8]For example, the LBS company GroundTruth described, as case studies on their website, how clients use their services to measure the impact advertisements have on driving store visits [29].

[9]As an alternative to publishing one static dataset, a trained model could be published instead. This would enable the generation of additional datasets that are also based on the real LBS data used for training.

Chapter 3

Generating synthetic mobility data with conditional neural networks: An implementation and evaluation framework to preserve utility and privacy

In the previous chapter I presented motivation for developing a system that can generate realistic, yet synthetic, mobility data. Namely, while location data collected from smartphones can be highly useful, it also has limitations such as small sample sizes and its sensitive nature. My contribution[1] in this chapter is to address these limitations with a new approach for generating synthetic mobility data that uses deep recurrent neural networks. The system I present is designed to take information about a population distribution as input in order to generate mobility data for that population. I also develop a framework to evaluate both the utility and privacy of the generated dataset. This includes a contribution to the computational privacy literature, with new "indistinguishability" criteria to show that the generated mobility data differ as much from the real data as the real data differs from itself.

[1]While the described work is my own, it benefited from invaluable ideas and feedback from Ronan Doorley and Esteban Moro.

The data generated by the models in this work retain aggregate statistical properties of the real data, as well as the patterns present in mobility traces at the individual level. The data also sufficiently varies at the individual level in order to protect user privacy. This work uses the location based services dataset referenced in the previous chapter in order to train the models and evaluate the output.

3.1 Introduction

Location data collected from user devices represent the histories of the device users. These location histories represent the behaviors of the individuals from whom it is collected, showing where they go and when, and it is laden with patterns that illustrate the routines of everyday life. Additional information can be inferred for the users represented in the data, such as home and work locations, as well as demographics.

As the previous chapters described, these data are highly valuable to the parties who collect it and those who pay to use it, and it is also highly useful to researchers and departments of transportation (DOTs) and other public agencies [30, 22].

However these data have limitations and their usage presents issues. One limitation is size. Obtained datasets often only represent a small sample of the population, making them less useful for analyses that are meant to address the full population. Another issue is privacy. Location data can reveal sensitive information about the people whose locations were collected, such as where they live, work, and frequent, and other places they went and when they went there.

A simple approach to protect user privacy is de-identification by means of attaching random identifiers to user data and removing identifying attributes of individuals. However, this is insufficient for spatiotemporal data where a subset of data points or areas visited can be unique to a user. Such a subset of data points is referred to as a location-based quasi-identifier (LBQID) [31] and researchers have demonstrated that knowledge of only a few spatiotemporal points or areas are needed to form an LBQID and re-identify most users in a de-identified dataset [32]. Researchers have also shown a variety of ways to use "side information" that is collected from other data sources, to re-identify users in location

datasets (in what are known as record linkage attacks) [33, 34, 35, 9].

Prior works have attempted to mitigate these privacy attacks with techniques known as "cloaking", such as data suppression [36], adding noise to location datasets [37], or otherwise reducing data accuracy or precision [38, 39]. Other techniques include segmenting single mobility traces into multiple pieces with different pseudo-identifiers [40] or swapping points between users' mobility traces [41]. Yet even so, researchers have shown that the privacy and re-identification risks that these techniques attempt to mitigate are still present [9]. Moreover, by modifying these datasets to mitigate privacy risks, these techniques decrease the utility of the data. This trade-off between privacy and utility is a well-acknowledged and studied problem inherent in the publication of microdata such as the location datasets addressed in this work.

In this work we approach the problem of the utility and privacy trade-off, as well as the problem of small data samples, by developing a machine learning model that generates synthetic data. We use a location based services dataset collected from user devices in 3 counties surrounding Boston, Massachusetts, and that represents roughly 2.7% of the area's population, in order to train the model and evaluate its output. The mobility traces used are from individuals over a 5-day workweek and the synthetic mobility traces generated by the model are representative of the real users' activity over that same time period.

The model is designed to generate a realistic synthetic dataset that retains the utility of the real data by retaining its properties, while mitigating privacy risks because it represents synthetic users whose data sufficiently differs from real users' data at the individual level.

Furthermore, our model is a conditional model that uses home and work areas as labels and as inputs in order to generate realistic location data for synthetic users with the given home and work areas. While synthetic data addresses the issue of privacy, this approach also addresses the issue of limited sample sizes. Population data, such as that reported by the census, can then be used to create the input data necessary to output a synthetic location dataset that represents the true population in size and population distribution.

Other works have also addressed privacy issues surrounding trajectory microdata with the generation and use of synthetic data. However we take a new approach in how we use deep

recurrent neural networks that have been successful in text generation as conditional models to generate synthetic mobility traces.

Contribution. Our approach exploits the patterns inherent in individuals' mobility traces in order to generate realistic synthetic mobility traces that retain these patterns. At the same time, our approach allows calibrating the amount of random noise in the model to better manage the utility and privacy trade-off. Calibrating the randomness enables our system to balance the extent to which generated mobility traces retain the properties of the real dataset (utility) with the extent to which they vary at the individual level (privacy).

Our contributions are further described in the context of related works in later sections. In short, other works do not address the full scope of this work, which we summarize as follows.

Our system generates realistic spatiotemporal data that represents users activities over an extended duration of time. The system takes home and work locations as inputs, enabling it to generate data for a given population distribution.

We also develop an evaluation framework to address the utility and privacy of synthetic trajectory data that is more comprehensive than other works. This includes new "indistinguishability" privacy criteria to establish that the synthetic dataset differs as much from the real dataset as the real dataset differs from itself.

Outline. This chapter first discusses related approaches to the generation of synthetic mobility data and how this work differs. It then describes how we model the problem of synthetic mobility data generation, including the description of our deep neural network model. It then provides our evaluation framework in the context of how others have approached location data synthesis and anonymization, in terms of both utility and privacy. Our evaluation framework is divided into separate sections for utility and privacy. Each section presents an overview of related concepts and works, before presenting our framework and metrics. Finally, we present our implementation with real data and our recurrent neural network (RNN) model, and evaluate the resulting synthetic data that it generates with the described evaluation framework.

3.2 Related Work

There is a large body of work using location data collected by mobile devices to extract information about a population's trips between places. Many of these previous works use call detail records (CDRs) to construct origin-destination matrices [42, 43, 28], as well infer users' home and work locations from this data for use in a variety of applications [44, 45].

Location data provided by CDRs is similar to location based services data collected from smartphone applications (which is what our case study uses) in that the location data is passively collected from mobile devices. However the location information from location based services data is more precise. Our work is applied to location based services data and builds upon methodology previously applied to CDRs.

Many works with CDRs address the problem of limited sample sizes that we named above by labeling individuals' data with inferred home and work locations and then using census data to expand their datasets to match known population quantities [22, 46]. A common approach to such mobility data generation processes is to derive aggregate statistical properties from the real data and then use these properties, often as parameters, in generative algorithms to produce synthetic trajectories for individuals. These processes are often implemented as markov chain models [47, 46].

Many of these generative algorithms are designed as "Exploration and Preferential Return" (EPR) models [48, 47, 49], where exploration is a random walk process [48] and preferential return accounts for the likelihood of people returning to previously frequented locations [50]. They leverage the predictable nature of human mobility and are designed to have users at their predetermined places of home and work during predefined hours for home and work, respectively.

However, the aforementioned works tend to focus on only the utility of datasets, and maintaining or enhancing that utility in the generation of synthetic data, without addressing privacy. More recently, there are also generative algorithms designed to balance the trade-off between utility and privacy for synthetic location trajectories by using differential privacy mechanisms [51] (differential privacy is further described in the privacy evaluation section). This includes the n-gram model by Chen et al. [52, 53] and the DP-Star [54], DPT [55], and

DP-WHERE [56] projects. They operate under a predefined privacy budget (ε) and spread that budget throughout their data generation process. They do this either by injecting calibrated levels of laplacian noise in each step of their generation process, or by only keeping elements in their generation process that maintain the conditions of the predefined privacy budget. There are many more works that similarly employ differential privacy for synthetic data generation, but with methods designed for relational databases and image datasets and that cannot be readily applied to sequential location data [57, 58, 59, 60, 61, 62].

The above cited works that generate synthetic mobility data with differential privacy claim to achieve privacy due to their generation process. However, they do not evaluate privacy by inspecting the generated output. Their work remains theoretical and they do not address whether the generated data is "private enough" or what their process of generating data with noise means for the privacy of users in the original datasets used for their generation processes. In our evaluation framework, we evaluate privacy with metrics that compare the generated output to real data.

Other issues with the above mentioned generative algorithms include their inability to simultaneously capture individual level patterns while also allowing users to break away from well defined patterns, as well as capture global patterns across individuals.

We also consider works that use neural networks to address problems related to the generation of sequential mobility data. These include generating non-sequential origin-destination mobility data [63] and traffic forecasting [64, 65]. Other related works use neural networks to generate time series, such as stocks data, but where the values in the generated data lack the geospatial information and relationships of mobility data [66, 67]. Kulkarni et al. (2017) [68] use recurrent neural networks to generate sequential traffic data, which we find most similar to our work. However, their model does not perform well by their own utility metrics and they do not address the evaluation of privacy.

Moreover, none of these related works can fully address the goal of this work which is to simultaneously do the following.

1. Generate realistic spatiotemporal data representing users' activity over an extended duration of time and where the generated data matches a desired population distribu-

tion, such as a distribution of home locations.

2. Balance the utility *and* privacy of the generated dataset, where utility is evaluated by how well the generated data retain properties of the real data.

Our contribution is to meet this goal by building upon these previous contributions. We further discuss in more detail the related work in utility and privacy evaluations for synthetic data, and how we draw from this work, in our evaluation framework sections.

3.3 Modeling the problem

This section describes our approach and methods. The details of our implementation are in section 3.6.

The goal of this work is to design a system that when given home and work locations as inputs, the system produces realistic location data for users with those home and work locations, where the location data represents an extended time period spanning multiple days or weeks. In this case, realistic is determined by how well the output corresponds to the real data, as evaluated by the evaluation framework. At the same time, the system should introduce variability. When given the same home and work location pair multiple times, the system should be able to produce different data each time, and the produced data must also sufficiently vary from the real data in order to protect the privacy of users whose data is in the real dataset. This too is evaluated by the evaluation framework.

We use home and work locations as inputs because data sources such as the census provide demographic information on where people live and work. This information can then be used by a system such as ours to produce a synthetic population with demographics representing the true population. Even in the case when home location may seem sufficient for representing the desired population demographics, the work location can play an important role in generating a synthetic population with sufficient variation, where the users from a single home area exhibit different mobility patterns.

3.3.1 Primitive functions

In this work we define and use two functions as primitives.

$$inferHome(location\ data) \rightarrow likely\ home \tag{1}$$

$$inferWork(location\ data) \rightarrow likely\ work \tag{2}$$

The $inferHome$ function takes a single user's location dataset as input and returns a likely home location, determined by where they spent the most time in the nighttime hours. Likewise, the $inferWork$ function returns where a user spent the most time during the workday hours (Monday to Friday). We call this the work location, but it could instead represent a different type of secondary location to the home, or might be the same as the home location.

These functions are used throughout this work for multiple purposes, including the labeling of training data as well as the evaluation of output data. In our implementation with a real geolocation dataset provided by a location based services company, we also use these functions to evaluate how well this data represents the true population estimates reported by census data (see section 3.6).

3.3.2 Mobility patterns

Our model for synthetic data generation is informed by how geolocation and mobility datasets are laden with patterns that reflect the routines of everyday life [69, 70]. For example, people tend to spend the night time hours in their home, and the hours of a work day at their place of work. Additional places people frequent may reflect other routines in people's lives, such as dropping children off at school, going to an art class, or grocery shopping. We can see these patterns in the real data by plotting the places an individual frequents over the hours of the day (see figure 3-1). These patterns should also be present in the generated synthetic data.

Our methods exploit these patterns. We consider a user's location data as a sequence of

device ID	latitude	longitude	timestamp	dwelltime
abc1234xyz345	42.472539	-71.107958	2018-05-06-18:11:1	5.02
abc123xyz345	42.427205	-71.014071	2018-05-06-19:01:53	45.10
def456qrs678	42.485207	-71.172924	2018-05-07-03:17:38	2.03

Table 3.1: Fake set of user location data, representing rows of a real LBS dataset. Each latitude and longitude point is a geolocation recorded by the associated device ID at the given timestamp. "Dwelltime" represents how long the user stayed in the reported location.

t_1	t_2	t_3	t_4	t_5	...	t_T
A	B	B	null	C	...	D

home	work	t_1	t_2	t_3	t_4	t_5	...	t_T
A	C	A	B	B	null	C	...	D

Table 3.2: Top: A representation of a "stay trajectory". Bottom: The same stay trajectory prefixed by a home and work location. Areas visited are represented by letters.

places visited and transform our data and choose our neural net architecture accordingly.

3.3.3 Data representation

Location based services (LBS) datasets are often collected and provided as timestamped geolocations, with latitude and longitude coordinates, where each data point has an associated pseudo-anonymized device ID for the user device from which the data was collected. It may also include how long the device reported a user stayed in the reported location. An example is provided in table 3.1.

We transform this dataset into a set of what we call "stay trajectories" for each user. An example stay trajectory is shown in table 3.2.

The indices of each user's stay trajectory (t_1, t_2, ..., t_T) represent time intervals, while the values represent the places, or areas, where the user stayed for the most time within the associated time interval. Areas are often repeated across time intervals, or sometimes the area is null valued in time intervals when no location data was reported for the user's device[2]. In this work we use census statistical areas to represent the places users stay because there is published demographic information about how many people live and work in these geographies. Census statistical areas are available in varying levels of granularity such as

[2]While we could infer values for missing data, our work is focused on the use case of generating synthetic data with properties similar to the true original dataset, including its sparsity.

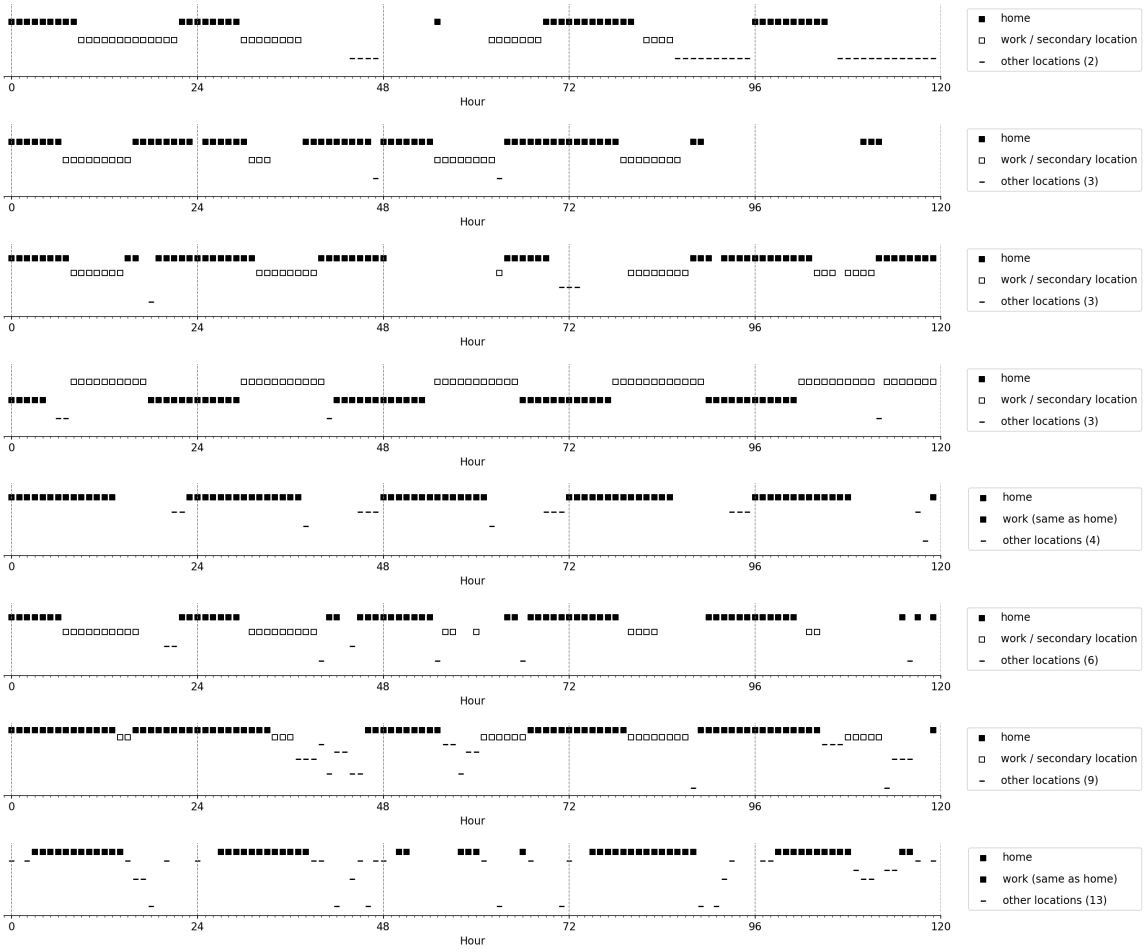


Figure 3-1: Mobility patterns seen in a sample of our user location data. Each plot represents a single user's sequence of places that they visited over a period of a 5-day workweek. The period is divided into 1-hour intervals. The hour is on the x-axis. A point is plotted above each hour of reported data for the user where the point represents the place the user spent the most time within that hour. When no data is reported for a user within an hour then no point is plotted. The y-axis represents the total amount of time the user spent in a given location relative to the other locations visited over the 5-day period: Locations were sorted by the total number of hours the user spent there, and the y-axis value indicates the sorted order. Distinct locations that were visited the same number of times are plotted with the same y-value. Home and work/secondary locations and any other locations are each plotted with different icons. These plots show patterns and user tendencies to either return to frequented locations or visit new ones at similar hours of the day, as well as how individual users' temporal patterns and routines vary across users. More plots can be viewed in our evaluation code notebook: `https://github.com/aberke/lbs-data/blob/master/trajectory_synthesis/evaluation/evaluate_rnn.ipynb`

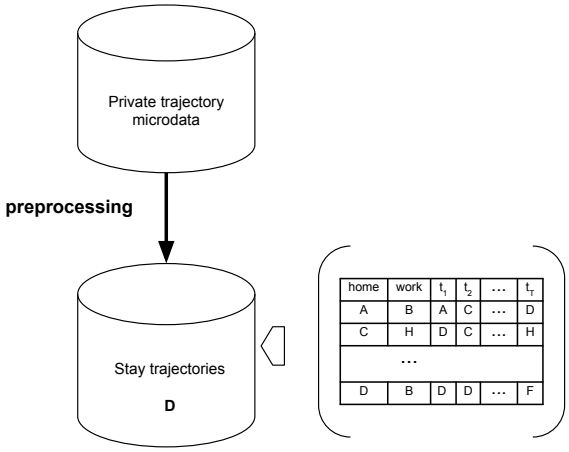


Figure 3-2: The diagram represents our preprocessing steps to transform tabular LBS data into "stay trajectories" that are then prefixed by their inferred home and work locations.

census tract, census block group, or census block. This level of granularity can be considered a tunable parameter. Likewise, the size of the intervals in which a time period is divided is tunable as well. With enough data, finer levels of granularity can be chosen for more precision.

Stay trajectories are thus sequences of areas representing where a user stayed, where the areas are discrete values. The area values have a spatial relationship, as some are spatially close together and therefore more likely to appear close together in a sequence, while others are miles apart, and therefore highly unlikely or impossible to directly follow one another in a stay trajectory.

There are further relationships between these areas and the patterns in which they appear embedded in these sequences. For example, someone who lives in area A, and works in area B is likely to spend many hours during the week day in area B and return to area A each night. Or people who spend time in area C may also tend to spend time in area D, or people may not spend time in area E during nighttime hours.

The ideal model learns the relationships between areas and the patterns and distributions of how people spend their time in these areas. The model should generate realistic sequences of areas based on these patterns at the individual level, while also introducing variability. The generated data should also retain aggregate statistical properties of the real training data.

3.3.4 Data definitions and model inputs and outputs

D: The full dataset of real stay trajectories.

S: A sampled subset of D.

S': A set of synthetic stay trajectories generated by the model.

$s = \langle s_1,\ s_2,\ ..., s_T \rangle\ \epsilon\ S$ represents a stay trajectory for a user, where each s_i represents an area. A $\langle home, work \rangle$ pair is associated with each s in S, where the home and work values represent areas and are also s_i values in s.

The $s'\ \epsilon\ S'$ match the format of s and represent stay trajectories for synthetic users over the same time period.

The model is trained with the entire dataset D. S is randomly sampled from D and used to generate input for the model and to then evaluate the output.

Each $\langle home, work \rangle$ pair for each sampled $s\ \epsilon\ S$ is used as input for the model to then generate a corresponding s' with the same $\langle home, work \rangle$ pair. This results in s' such that $|S| = |S'|$ and where the distribution of $\langle home, work \rangle$ pairs is consistent for S and S'. These consistencies are crucial for proper utility and privacy evaluation, as the evaluation compares the real data to the synthetic data. However, any distribution or number of $\langle home, work \rangle$ pairs can then be used as input for the model to generate a synthetic dataset.

3.3.5 Deep recurrent neural network model

Text and music have similar properties and patterns as stay trajectories. There are temporal and spatial relationships between words in text and notes in music.

In each case, data can be considered as a sequence of tokens, where the tokens can be characters or words in text, notes in music, or areas in stay trajectories. In this way we consider stay trajectories analogous to sentences of text or lines of music, where each area in a stay trajectory sequence is analogous to a word or musical note.

Recurrent neural networks (RNNs) that use long short-term memory (LSTM) [71] units have

been successful in generating complex sequences [72] that retain the structural properties inherent in text [73, 74] and music [75, 76, 77], which we also see with our stay trajectory data.

RNNs are trained by processing each sequence in a training set one element at a time, and predicting a next element. Each prediction is conditioned on the previous elements encountered in the sequence. That is, for a generic input vector sequence $x = \langle x_1, ..., x_T \rangle$ and output vector sequence, $y = \langle y_1, ..., y_T \rangle$, each output vector y_t is used to parameterize a predictive distribution, $Pr(x_t + 1|y_t)$. The loss of wrong predictions is propagated back through the network for the model to "learn" from.

This same process can then be used for sequence generation by feeding the model's predictions back to the model as input for the next step as if they were real rather than the model's own inventions. Each prediction step injects stochasticity, where the model samples from a distribution of candidate next elements that is conditioned on the previous elements. This sampling process offers the opportunity to introduce varying levels of randomness and variation in the output. The overall process allows for the generation of novel sequences that are similar to the training set. It also simulates a high-dimensional interpolation between training examples that distinguish RNNs from n-gram or other generative algorithms.

RNNs can be further improved with the Attention mechanism [78]. With Attention, a vector of importance weights can be learned by the network and then used to predict a next element in a sequence based on how strongly it is correlated to, or "attends to", previously encountered elements.

As described above, RNNs are conditional models by nature in that they predict a next element in a sequence conditioned on previous sequence elements. We leverage this conditional process for our use case.

We concatenate each stay trajectory, s, in our training set with its associated $\langle home, work \rangle$ label pair, resulting in a sequence of tokens where the $\langle home, work \rangle$ label pair is a prefix for s, $\langle home, work : s \rangle = \langle home, work, s_1, s_2, ..., s_T \rangle$. These prefixed stay trajectories are used to train the model. The model learns the relationships and patterns of tokens (areas) in the sequences. It also learns the relationship between the $\langle home, work \rangle$ area prefixes

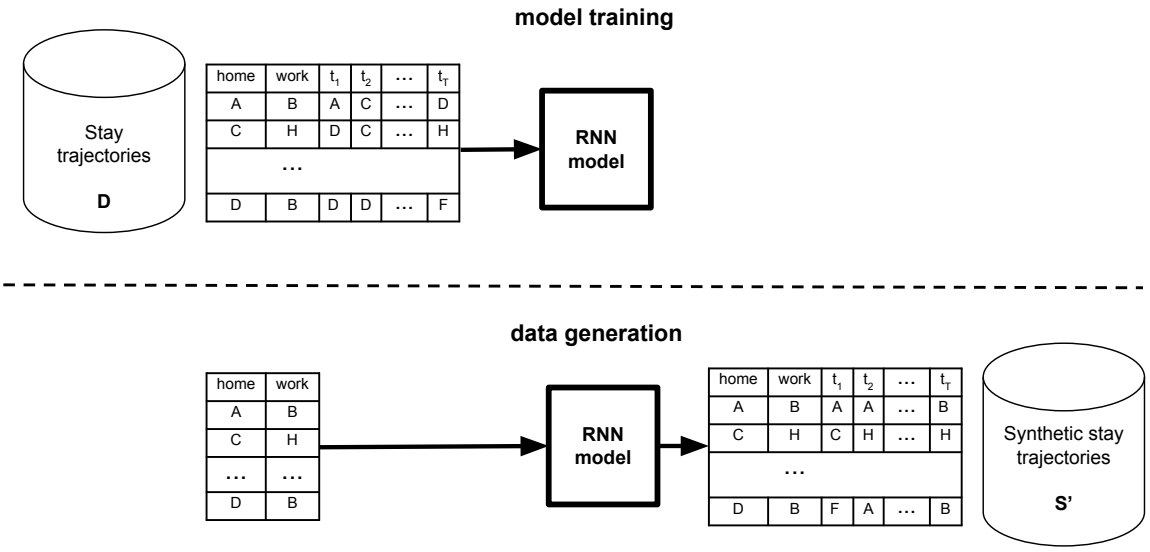


Figure 3-3: The model is trained with real location trajectory data where each user's trajectory data is labeled by home and work locations. The model is then used by taking home and work location labels as input to generate corresponding synthetic trajectory data.

and the distribution of area tokens that then follow in s. For example, it might learn that the area token in the home prefix position is most likely to occur in the sequence positions representing nighttime hours, and similarly that the area in the work prefix position is a more likely candidate for sequence positions representing work hours, as well as learn other structural relationships and patterns between area tokens.

For the generation phase, we then feed $\langle home, work \rangle$ label pairs to the trained network for it to generate corresponding stay trajectories. The network treats the input $\langle home, work \rangle$ pairs as prefixes for sequences that it has learned to complete. The sequences of generated tokens that follow are the generated synthetic stay trajectories, s' in S', with the given $\langle home, work \rangle$ label pairs.

Basic deep recurrent neural network. A recurrent neural network (RNN), such as the model used in this work, is a network of "neural" unit nodes organized into layers. An exterior "input" layer of nodes receives input, and an exterior "output" layer of nodes produces the network's output. Between these are "hidden" layers of nodes where additional layers add additional depth to a deep neural network. Each node within a layer has a one-way weighted connection to each node in the successive layer. The weights for the connections are

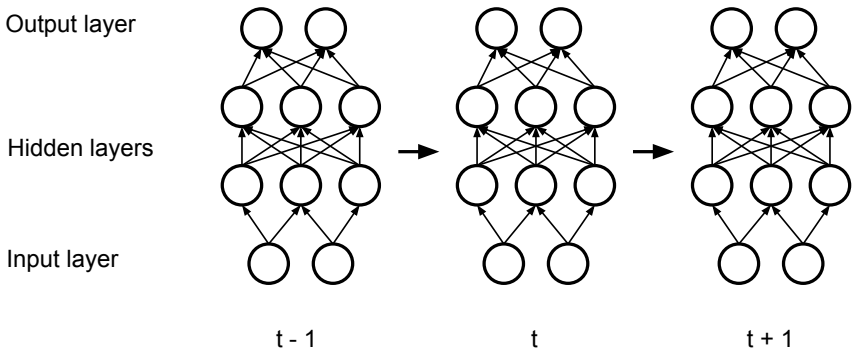


Figure 3-4: A recurrent neural network with 2 hidden layers, adapted from figure 1 of Sutskever et al. (2011) [74]. Weights between units are shared across time. The nonlinear activation function used by the hidden units is a source of the RNN's rich dynamics.

learned during the training process, but the architecture for the network, such as the number of layers and the type and number of neural units within each layer, must be determined beforehand. The parameters that determine such variations to the network architecture are "hyperparameters".

Each node in the network has a nonlinear "activation function" that allows the network to share varying levels of state across time as it sequentially processes items, while also providing a source of rich dynamics. A basic illustration of an RNN is shown in figure 3-4.

Network architecture and hyperparameters. The recurrent neural network architecture used has a series of connected layers illustrated in figure 3-5. The embedding layer encodes the input into a form for the network to process. It is followed by "hidden" layers of LSTM units. The LSTM is bidirectional to improve the training process. The embedding and LSTM layers are each skip-connected to the attention layer, which is connected to the final output layer[3].

[3]The system borrows from the textgenrnn project architecture, `https://github.com/minimaxir/textgenrnn`.

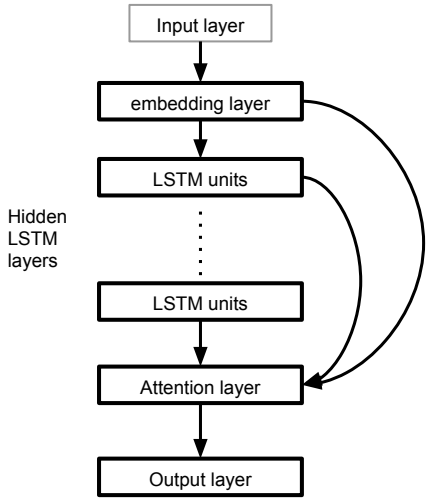


Figure 3-5: The recurrent neural network is a series of connected layers. An embedding layer encodes the input into a form for the network to process. It is followed by "hidden" layers of LSTM units. The embedding and LSTM layers are each skip-connected to the attention layer, which is connected to the final output layer.

This general architecture has a variety of hyperparameters.

- Dimension of the embedding: The embedding determines how each token in a sequence is represented for the network, and is learned during training. The dimension of the embedding is a hyperparameter.
- Layer size: The number of LSTM units in each hidden layer.
- Layers: The number of hidden layers.
- Dropout: The rate at which weighted connections between neural units are randomly excluded for each training sample. Dropout is a regularization method that helps prevent overfitting and improves model performance.
- Maximum length: The maximum number of input tokens in the sequence that the network considers when predicting the next token. This number should be long enough to allow the model to learn recurrent patterns in sequences, yet additional length adds complexity.
- Temperature: The amount of random noise added to the predictive sampling for each next token in a sequence.

In our implementation (see section 3.6) we tested models with a variety of these hyperparameters and their combinations and evaluated results with our evaluation framework.

3.4 Evaluation framework: Utility

3.4.1 Related work in utility evaluation

Works that apply differentially private techniques are helpful references, although they often use relational data rather than sequential data. They commonly evaluate statistical utility by computing the distribution of attributes in their datasets, or the distribution of k-way marginals drawn across attributes (for some small k), and comparing the distributions to those computed for the original dataset [57, 79]. Total variation distance [80] is a metric used to quantify the difference in distributions. Another evaluation metric used is the agreement rate in a machine learning label prediction task [57, 79]. Specifically the agreement rate is defined as the percentage of records for which two classifiers make the same prediction, where one classifier uses the real original data, and the other classifier uses the sanitized data.

Another evaluation strategy uses counting queries. This is more common with location trajectory data [54, 55, 81] and other sequential data generation or sanitization frameworks [53, 82]. A counting query defines $Q(D)$ as the count of items in dataset D that satisfy the query Q. For example, Torres et al. (2016) [81] use spatiotemporal range queries to count the number of users within a given area within a given time period. They then evaluate sanitized trajectory databases by how closely the query counts match the query counts returned by querying the original data. A common way to quantify this evaluation uses relative error, $RE = \frac{|Q(D)-Q(S_D)|}{max\{Q(D),b\}}$, where S_D is a sanitized dataset that is meant to mimic the utility of D, and b is a sanity bound.

Differentially private sequential data generation frameworks such as DP-Star [54], DPT [55], and the n-gram model by Chen et al. (2012) [53] also use sequential pattern mining as an evaluation strategy. They identify the most frequent sequential patterns, where a pattern is a subsequence of a full trajectory sequence. Utility loss is then quantified with the rate at

which the most frequent patterns change between the original and sanitized datasets. The DPT and DP-Star projects also evaluate utility with a measurement over the distribution of where trips originate and end. DPT measures trip distribution quantitatively, while DP-Star uses visuals to evaluate how well their framework preserves the spatial density of real taxi trips data. They plot the trip origin and destination points in a single color on an x-y plane for both the real data and their generated synthetic data to evaluate how well their system preserves the spatial density of these points (see figure 1 from Gursoy et al. (2019) [54][4]).

While these works provide helpful reference, we note that the form of the sequences they evaluate differs from our dataset. They either lack temporal information, or they represent moving objects over shorter durations, where adjacent items in a sequence likely differ. For example, n-gram uses sequences of station visits by Montreal transportation system users, and DPT and DP-Star use taxi trips data and trips data generated by Thomas Brinkhoff's network-based generator for moving objects [83]. Each of these trips is considered a full trajectory sequence rather than how our stay trajectories follow users across a time period with a series of days and trips, and where patterns should be captured at a higher level. Moreover, in our dataset we expect subsequences that represent users staying in a single location for a long duration, or subsequences with no reported data. Despite these differences, we make special note that these related works examine patterns, measure aggregate distributions of land use, and use visuals; our utility evaluations do this as well.

3.4.2 Utility evaluation used

Our utility evaluation framework is composed of multiple tests and metrics. We describe each one in this section with more implementation details and results in section 3.6.

- Home and work labels match generated data.
- Trips between places are realistic.
- The aggregate distribution of where users spend time is consistent with real data.

[4]Link to image for figure 1 from Gursoy et al. (2019) [54]: `https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/7755/8821494/8481494/gurso1-2874008-hires.gif`

- The distribution of the number of distinct places individuals visit within a time period is consistent with real data.

- Mobility patterns at the individual level are retained.

Home and work labels match generated data. Our model is designed to take $\langle home, work \rangle$ label pairs as input and output corresponding synthetic stay trajectories. We apply the $inferHome$ and $inferWork$ functions to the output synthetic stay trajectories and quantify the rate at which the input labels match the inferred labels. That is, for each input label pair $\langle home_i, work_i \rangle$ and corresponding synthetic trajectory s'_i in S' there is a home label match when $home_i = inferHome(s'_i)$ and a work label match when $work_i = inferWork(s'_i)$. We quantify the number of home and work label matches.

Trips between places are realistic. Synthetic stay trajectories should not contain any sequences of stays that represent impossibly far trips within the time period the trip spans, and trips users make in the synthetic data should be consistent with the real data.

We consider two consecutive stay locations in a stay trajectory as a bigram, (A, B). First we check that the geographic distance between all bigram locations can be traveled within the time period represented by the bigram. Second, we call a bigram (A, B) that occurs in the synthetic data an "unseen bigram" if neither it nor its reverse, (B, A), occur in the real data. We count the total number of "unseen bigrams", with duplicates, and quantify the portion of unseen bigrams as the total number of "unseen bigrams" over all total bigrams in the synthetic data, $\frac{(unseen\ bigrams\ in\ S')}{all\ bigrams\ in\ S'}$.

The aggregate distribution of where users spend time is consistent with real data. Individual stay trajectories vary in where the users they represent go, and when. However, in aggregate the distributions of where users spend time should be consistent across the real and synthetic data. We compute the aggregate amount of time intervals users spend in each area for the real data, S, and synthetic data, S', and measure the correlation between their distributions. Note that where users spend time is biased to where they work and live, which is why it is important to compare S' to S (rather than D), where the distribution of

$\langle home, work \rangle$ label pairs is consistent across the two datasets.

The distribution of the number of distinct places individuals visit within a time period is consistent with real data. There is variation in the number of distinct areas in each stay trajectory due to the heterogeneity in users' levels of activity or activity diversities. For example, some users have stays in only 2 or 3 distinct areas per week, while other users have stays in many more distinct areas, either because they have more active lives and visit more different places, or because their devices report data more often. This distribution of distinct places per stay trajectory should be maintained across real and synthetic datasets.

We use the Pearson's chi-squared test for homogeneity to determine whether this distribution is maintained by considering S and S' two samples that might be drawn from the same population. Specifically, we consider the number of distinct areas in each stay trajectory as a category and count the frequency of this category in each of S and S'. We then test the null hypothesis: The proportion of s' in S' with P distinct areas is the same as the proportion of s in S with P distinct areas, for each P occurring in either S' or S.

We test the null hypothesis with a significance level of 0.05.

Mobility patterns at the individual level are retained. Similar to the DP-Star authors [54], we use a visual mechanism to evaluate the quality of our synthetic data. We visualize individual mobility patterns over the duration of stay trajectories by plotting visited areas over the intervals of stay trajectories, and comparing plots from the real and synthetic datasets (see figure 3-1).

3.5 Evaluation Framework: Privacy

In order to validate the value of using synthetic data as a means to protect user privacy, it is necessary to evaluate how well a synthetic dataset, which is based on a real dataset, avoids leaking private information about individuals represented in the real dataset. As an extreme example, if systems such as ours are overtrained, or do not generate data with sufficient randomness, then generated synthetic trajectories might match real trajectories

from the training set (in which case a system might just as well sample from the real dataset). With the specifics of our machine learning system in mind, we approach privacy by evaluating whether individual stay trajectories sufficiently differ between the real and synthetic datasets.

3.5.1 Related work in privacy evaluation

There are well established privacy criteria that are related to our work. They tend to measure "indistinguishability" as a way to validate that any private record in a database is not distinguishable from a large enough group of other records. This section describes them, as well as their limitations with respect to our use case. We build upon them to establish our method for privacy evaluation.

k-anonymity. k-anonymity was first published as a privacy criterion for relational micro-data [84] and has since been adapted and widely used for spatiotemporal trajectory data. The central concept is that a subset of points in any user's data should be present in $k-1$ other users' data. When this is the case then k-anonymity is achieved, since k users are indistinguishable. When applied to spatiotemporal trajectory data, a subset of a user's geolocation points is considered a location-based quasi-identifier (LBQID) and k-anonymity can be formally defined as follows [9].

Definition for k-anonimity. Let D be a database of trajectories and $LBQID$ the associated location-based quasi-identifier, and let $D[LBQID]$ be the set of records returned by a query for $LBQID$ on D. Then, D satisfies k-anonymity if and only if there are at least k records in $D[LBQID]$.

Other works have achieved k-anonymity for spatiotemporal data via data suppression or other modifications [85, 86, 87, 88, 81]. However this definition of k-anonymity is not directly applicable to privacy risks for synthetic data. k-anonymity addresses the risk of a user being re-identified in a de-identified database based on the uniqueness of their data. But for synthetic data, there is no real user to be re-identified.

Differential privacy. Much of the modern literature about publication of private datasets focuses on differential privacy [51], which has been adopted in both academia and industry[5]. A core concept of differential privacy (DP) is in line with this work: Querying a published dataset reveals information about a population without revealing information about particular individuals. With ε-differential privacy there is a privacy "budget" parameter, ε , which regulates the amount of variation to expect from a query result when any individual in a dataset is removed from that dataset. In other words, if two datasets D_1, D_2, differ in only one sample, then the amount by which their results for the same query differ should be within the bounds defined with ε, in which case the querying mechanism is ε-indistinguishable.

Definition for ε-differential privacy. A randomized algorithm $A : D \rightarrow R$ satisfies ε-differential privacy if for any two adjacent databases, D_1, D_2, which differ in only one sample, and for any subset of output $S \epsilon R$, $Pr[A(D_1) \epsilon S] \leq e^{\varepsilon} \times Pr[A(D_2) \epsilon S]$.

Differential privacy is most commonly used to address privacy for relational databases, where records with many attributes often have only one or a few sensitive attributes that should be considered private. This kind of database differs from the trajectory microdata addressed in this work, where user records represent sequences of spatiotemporal data points. Moreover, differential privacy was developed for the purposes of data mining rather than synthetic data generation. Even so, differential privacy has been adapted and applied to work in differential private data generation.

ε-differential privacy is often achieved by applying Laplacian noise to the output of queries, where the noise is calibrated according to ε. There are compositional properties of ε-differential privacy that allow a total privacy budget, ε, to be divided among queries or other processes, so that when combined they can achieve more complex ε-differentially private algorithms. This has led to theoretically provable ε-differentially private data generation techniques [55, 62, 58, 60, 61, 57, 92, 93, 59, 56]. For example, deep neural networks have been designed to satisfy ε-differential privacy by injecting noise throughout their training processes, where the cumulative noise is tracked (often with a "moment accountant" [94]) and calibrated to the privacy budget, ε. Other works divide ε among sequential queries to

[5]Various implementations of differential privacy have been adopted by industry giants such as Apple, Google (Chrome), Microsoft (Windows 10 operating system), Uber, as well as by federal agencies such as the U.S. Census Bureau [89, 90, 91].

produce sequential data such as trajectory microdata. In these cases, the authors provide theory to prove that the algorithmic *process* of the data generation satisfies ε-differential privacy. However, they do not evaluate the privacy properties of the *output*, leaving comparisons between generated and real datasets, and therefore real privacy implications, outside the scope of their work.

Plausible deniability. Another privacy criterion developed to more directly address privacy risks and indistinguishability for synthetic trajectory data is "plausible deniability" [95]. It is developed for a model where a subset of seed records, such as S, is sampled from a dataset of real records, such as D, and where each seed is then transformed to produce a synthetic record. The plausible deniability criterion is then met if each synthetic record could have been generated by a sufficiently large number of real records, making the input for the generation process indistinguishable. It can be formally summarized for trajectory data as follows.

Definition for plausible deniability. A synthetic trajectory s_i' generated from a seed trajectory $s_i \, \epsilon \, S \subset D$ satisfies (k, δ)-plausible deniability if there are at least $k \geq 1$ alternative trajectories $s_j \, \epsilon \, D$ such that the similarity, σ, of s_i' and s_i is within δ of the same similarity measured between s_i' and any s_j, i.e. $|\sigma(s_i, s_i') - \sigma(s_j, s_i')| \leq \delta$.

In this definition, k is the threshold number of trajectories in the real database D with which any seed in s is indistinguishable, allowing the "plausible deniability" that s was used to produce s'. It requires defining a metric, σ, to measure the similarity between trajectories, as well as defining a threshold δ for similarity.

A limitation of the plausible deniability criterion is that it assumes each synthetic output record is produced by transforming a single real input record. This prevents it from being applied to more general data generation systems, including the system presented in this work - our system uses all real input records as training data input so that output synthetic records are produced as an interpolation across them.

A further limitation of each of the criteria above is their abstract nature; their real application requires determining reasonable privacy parameters, such as k and δ for (k, δ)-plausible

deniability, ε for ε-differential privacy, and k for k-anonymity. Determining these parameters is often left outside the scope of work.

Our evaluation method considers privacy and indistinguishability with attention to privacy parameters by measuring the indistinguishability of synthetic records among real records as compared to the indistinguishability of real records among themselves.

3.5.2 Privacy evaluation used

A goal of this work is to ensure any synthetic stay trajectory generated by a model, M, is sufficiently different from any real stay trajectory that was in the model training set.

For any two trajectories, s_i, and s_j, we use a distance metric to measure the difference between them, denoted $dist(s_i, s_j)$.

We compute the minimum distance between a given s and any other s_j in a set of stay trajectories, D, which we call $m\text{-}dist(s, D)$.

$m\text{-}dist(s, D) = dist(s, s_j)$ such that $\forall\ s_j, s_k\ \epsilon\ D, dist(s, s_j) \leq dist(s, s_k)$

Our privacy criterion then evaluates for a distance m, and small probability δ,

$Pr[m\text{-}dist(s, D) \leq m] \leq \delta$

The evaluation is over a probability to address the evaluation of the model M, which is a stochastic process used to generate S', rather than a specific static set S'.

The process of sampling real trajectories S from D (where the distribution of labels in S is then used to generate S' with a matching distribution) is also stochastic.

In order to determine the proper δ value for a given distance m and apply the criterion for a synthetic data generation model, we compute the distance metric for both the synthetic trajectories and the sampled real trajectories, and then compare their probability distributions.

Values $m\text{-}dist(s', D)$ are computed for each s' in S' as $dist(s', s_j)$ such that

$\forall\ s_j, s_k\ \epsilon\ D, dist(s', s_j) \leq dist(s', s_k)$.

Values $m\text{-}dist(s, D)$ are similarly computed for each s in $S \subset D$, but where direct comparison of s to itself is avoided.

The privacy criterion is then satisfied if, for any distance m,

$Pr[m\text{-}dist(s', D) \leq m] \leq Pr[m\text{-}dist(s, D) \leq m]$

Meaning that the probability that a synthetic trajectory s' differs from any real trajectory in D by less than m, is less than or equal to the probability that a real sampled trajectory, s, differs from other real trajectories in D by less than m. In other words, the level of indistinguishability between synthetic trajectories and real trajectories is at least the level of indistinguishability within the set of real trajectories which the synthetic trajectories are meant to represent.

Distance metric. To compute the difference between stay trajectories, we use the Levenshtein edit distance which was developed as a metric for sequences [96]. The Levenshtein edit distance between two sequences is the minimum number of insertions, deletions, or substitutions necessary to transform one sequence into the other. In our case, the tokens subject to insertion, deletion, or substitution, are the areas represented in stay trajectory sequences.

Prior works have used other difference metrics for location trajectory data, such as Euclidean distance [97][6], Longest Common Subsequence [98][7], Hausdorff distance [99, 100], Manhattan norm [81], and compared them [101, 102]. A comparison of similarity measures for location trajectory data by Chen et al. (2005) [102] found edit distance[8] to be a more robust metric in terms of accuracy and accounting for noise.

The Levenshtein edit distance can be recursively defined as the length between two se-

[6]When used as a metric for location trajectories, Euclidean distance is defined as the average euclidean distance between corresponding points within a trajectory, where the length of trajectories are the same.

[7]When used as a metric for location trajectories, Longest Common Subsequence finds the alignment between two sequences that maximize the length of common subsequence.

[8]The authors, Chen et al. (2005), refer to the edit distance used in their paper as "Edit Distance on Real sequence" (EDR) [102]. It is based on Levenshtein's edit distance and modified to handle real-valued locations, as opposed to the discrete values in the sequences used in this work.

quences, a and b, with lengths $|a|$ and $|b|$, respectively, as $dist_{a,b}(|a|, |b|)$, where:

$$dist_{a,b}(i,j) = \begin{cases} max(i,j) & \text{if min(i, j) is 0} \\ min \begin{cases} dist(i-1,j)+1 \\ dist(i,j-1)+1 \\ dist(i-1,j-1)+1_{a_i \neq b_j} \end{cases} & \text{otherwise} \end{cases} \quad (3)$$

The term $1_{a_i \neq b_j}$ is equal to 0 when $a_i = b_i$ and 1 otherwise. At any indices i and j, $dist_{a,b}(i,j)$ is the distance between the first i and j tokens of a and b, respectively.

3.6 Implementation and results evaluation

3.6.1 Data source

This work used location based services (LBS) data provided by a location intelligence and measurement company. The data was provided as pseudo-anonymized GPS locations from users who opted-in to share their data anonymously through a GDPR-compliant framework. Researchers followed a strict contract with obligations to not share data beyond aggregate statistics, or to attempt to de-identify data.

3.6.2 Data panel and preprocessing

The trajectory microdata was provided as tables of "stays" data where each row includes a pseudo-anonymized device ID, latitude and longitude coordinates, timestamp, and estimated time the device was active in that location (see table 3.1).

Geography and time period. We used data reported by user devices in 3 counties surrounding Boston, Massachusetts (Middlesex, Norfolk, and Suffolk counties) and used data from the first 5 day workweek of May 2018.

Total device count: 83,827

	data points	unique days of data	unique nights of data
mean	10.013	2.783	2.352
std	12.654	1.605	1.395
minimum	1	1	1
25%	2	1	1
50%	5	2	2
75%	13	4	3
maximum	222	5	5

Table 3.3: Statistics for the data reported by each device over the 5-day workweek used, from 3 counties surrounding Boston MA.

Data filtering and panel[9]. We dropped all data points representing more than 24 hours spent in one location. This resulted in 839,368 data points reported by 83,827 unique user device IDs. This population represents roughly 2.7% of the total population for the 3 counties, based on the ACS 2018 estimates [103].

However, the reported data is highly sparse, with high levels of variation in the number of datapoints reported by each device, and the number of unique days and nights for which their data is reported. Table 3.3 shows statistics for the number of data points reported by each device as well the number of unique days and nights for which data is reported by each device.

We restricted the data used in this work to user devices reporting at least 3 unique days and 3 unique nights of data. The resulting data panel includes data from 22,707 user devices, representing roughly 0.726% of the population. See the appendix (A.1) for additional information about data representativeness.

Granularity. The 5-day time period of stay trajectories is divided into 1 hour time intervals. Census tracts are used as the area values. Each item in a stay trajectory sequence then represents the census tract where a user stayed most in the corresponding 1 hour interval. These parameters are chosen due to the size and sparsity of our dataset, however, with more data our methods can be applied with higher levels of spatial and temporal precision to generate synthetic data with more information about user activity throughout the day.

[9]The data preprocessing code used can be viewed in our open source repository: `https://github.com/aberke/lbs-data/blob/master/preprocess_filtering.ipynb`.

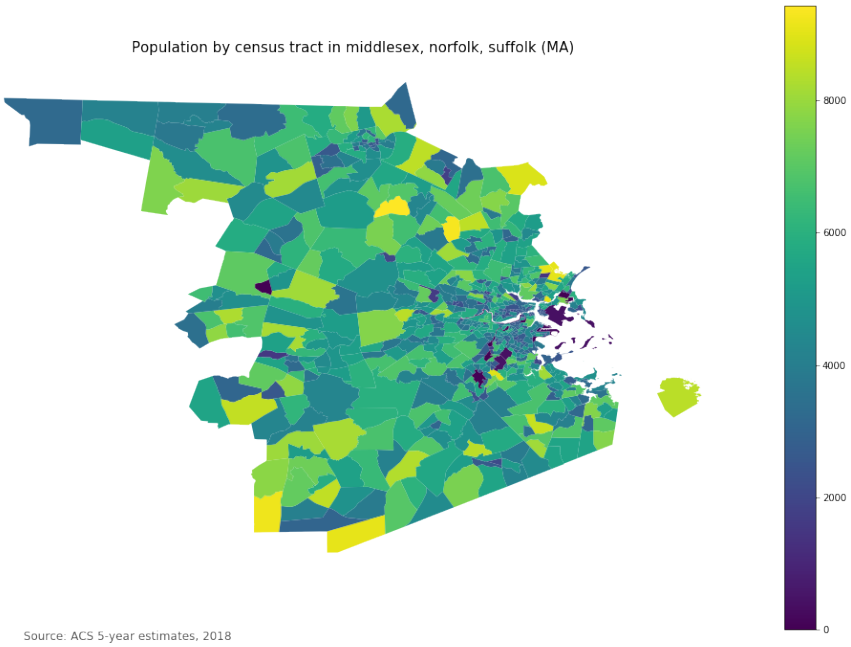


Figure 3-6: Population estimates from the 2018 ACS 5-Year estimates [103] for the 3 Boston area counties used for in our dataset. The provided location services data used in this work includes data from roughly 2.7% of the area population.

Data transformation. LBS data from the data panel was transformed into stay trajectories representing the location histories for each user device in the panel. Each stay trajectory is prefixed with its inferred home and work locations. The $inferHome$ and $inferWork$ functions are implemented with the nighttime hours that determine the inferred home locations defined as 8pm to 9am, while the remaining hours of the day are used to infer work location. We validated this choice of hours used for home inferences by comparing our resulting census tract population estimates to census data, and find a correlation of ρ=0.648 (see the Appendix section about data representativeness).

We note that as the resulting stay trajectories are used for the model training and generation processes, the area labels within them can be arbitrary. What is important for the model's success is the relationship between them. For this reason we map real census areas to integers, and map each area in stay trajectories to the integer representing the area. The transformed stay trajectories can then be more safely shared and used in remote computing environments while the mapping between real census areas and their integer representations

is kept private[10]. We use such transformed stay trajectories for model training and data generation, and then map the integers in the model's output stay trajectories back to the real areas they represent.

Data used for model training, generation and evaluation.

D is the dataset of 22,707 stay trajectories from panel users and is used to train the model.

S is a subset of 2000 stay trajectories randomly sampled from D.

S' is 2000 synthetic stay trajectories where the distribution of $\langle home, work \rangle$ label pairs is consistent between S and S'. S' is generated by providing the $\langle home_i, work_i \rangle$ label pair for each s_i in S to the trained model to produce a corresponding s'_i in S'.

Each stay trajectory has a length of 120 indices (5 days x 24 hours) and there is a token vocabulary size of 652 (corresponding to the census tracts in the chosen geography).

The code that implements what is described above can be viewed open source[11].

3.6.3 Model and evaluation

We trained over 70 models with a variety of hyperparameters and their combinations, as described in section 3.3.5. A summary of the hyperparameters and evaluation results for 8 of the best models is shown in table 3.4. These best models were chosen by evaluating their output. They were first filtered to meet a home match rate threshold of 0.85, then sorted and filtered by how well they met the privacy criteria. From the filtered list a "best" model, M, was chosen as the model that performed well by the variety of privacy criteria and utility metrics, and best by the chi-squared test for homogeneity where there was the

[10]We published the transformed stay trajectories used in this work to Github, where their real areas are mapped to arbitrary integers and the mapping between real areas and integers is kept private. `https://raw.githubusercontent.com/aberke/lbs-data/master/trajectory_synthesis/data/relabeled_trajectories_1_workweek.txt`.

[11]The code that implements what is described above is viewable open source. Code and tests for functions to transform and label data: `https://github.com/aberke/lbs-data/blob/master/trajectory_transformers.py`. Code notebook for transforming and prefixing the data panel and generating the samples used: `https://github.com/aberke/lbs-data/blob/master/trajectory_synthesis/trajectory_synthesis_notebook.ipynb`.

		Models							
		A	B	C	D	E	F	G	**M**
parameters	embedding dimension	100	100	100	128	128	128	128	**128**
	max length	72	72	72	70	70	60	60	**60**
	layers	2	2	2	3	3	3	3	**3**
	layer size	256	256	256	128	128	128	128	**128**
	dropout	0.3	0.3	0.3	0.1	0.1	0.1	0.1	**0.1**
metrics results	work label match rate	0.818	0.790	0.760	0.774	0.765	0.767	0.692	**0.733**
	portion unseen bigrams	0.006	0.014	0.036	0.002	0.005	0.003	0.015	**0.006**
	correlation for time in areas	0.942	0.934	0.936	0.915	0.916	0.935	0.930	**0.936**
	X^2 homogeneity test p-value	0.016	0	0	∼0	∼0	∼0	0	**0.429**
	minimum m-dist(s', D)	0	0	2	0	3	0	2	**2**
	1% cutoff for m-dist(s', D)	4	5	9	3	5.49	3	6	**5**
	5% cutoff for m-dist(s', D)	9	12	20	6	9	7	14	**10**
	10% cutoff for m-dist(s', D)	13	18	27	9	12	10.9	18	**14**
	10% cutoff for m-dist(s", S')	18	28	41	13	20	15	29	**20**

Table 3.4: Model parameters and results for 8 of the best models, where the models are sorted by home label match rate. The chosen "best" model is **M**.

most variation in results (this test was used to evaluate the distribution of the number of distinct places visited). In what follows, we describe **M** and describe the evaluation results for the synthetic dataset it generated, S'. All code for evaluations as well as results for additional models can be viewed in our open source code notebook via Github[12].

The chosen "best" model **M** has a 128-dimensional embedding layer, and is composed of 3 bidirectional LSTM layers with 128 LSTM units each and a 0.1 dropout.

3.6.4 Utility evaluation implementation and results

Home and work labels match generated data. We apply our $inferHome$ and $inferWork$ functions to each pair of input labels $\langle home_i, work_i \rangle$ and corresponding output s'_i in S' and quantify the home label matches where $home_i = inferHome(s'_i)$ and work label matches where $work_i = inferWork(s'_i)$. We measure the quantities as a portions over S',

home match rate $= \frac{\text{home label match}}{|S'|}$,

work match rate $= \frac{\text{work label matches}}{|S'|}$.

[12]Open source evaluation code and results for additional models are in the notebook: `https://github.com/aberke/lbs-data/blob/master/trajectory_synthesis/evaluation/evaluate_rnn.ipynb`.

What is a reasonable match rate? We cannot expect a match rate of 100%. Variation exists within the real data that we quantify and compare our results to. We compute a secondary user data panel, defined with the same criteria as described in section 3.6.2, and over the same geography, but for a different 5-day workweek. (The primary user panel used to produce D and S is for the first 5-day workweek of May 2018. The secondary user panel is for the second 5-day workweek of May 2018.) 62% of the users in the primary data panel (N=22673) also met the data reporting criteria for inclusion in the secondary user panel (14076 out of N=22522). For users in both data panels, we apply the $inferHome$ and inferWork functions to each week of their data separately and count a match as when the inferred label is consistent across the weeks. The home label match rate is 91.3%. The work label match rate is 75.4%. We use these quantities as benchmarks for our synthetic data evaluation[13].

The home label match rate for S' is 86.2% and the work label match rate is 73.2%. We note that other models performed better by this metric, with home and work label match rates above the real sample benchmarks (see table 3.4), however they did not perform as well in other respects, such as privacy.

Trips between places are realistic. The portion of unseen bigrams over S' is 0.5%. In other words, of all the trips between places in the synthetic data, only 0.5% do not also occur in the real data. We verify that the synthetic data does not have users make impossibly far trips. Since any bigram in the real data is a possible trip, we only check the "unseen bigrams". We measure distance between areas in a bigram as the line distance between the centroids of the census tracts they represent. See figure 3-7 for a histogram of distances between all "unseen bigrams" in S'. Given that two consecutive time intervals in our stay trajectories data represent a period of two hours, and users in our dataset can drive 50 miles per hour in a car, we verify that the 2 consecutive areas of all "unseen bigrams" in S' are within 100 miles of one another, which they are.

[13]Computation of these benchmarks can be found at: `https://github.com/aberke/lbs-data/blob/master/evaluate_home_work_changes.ipynb`.

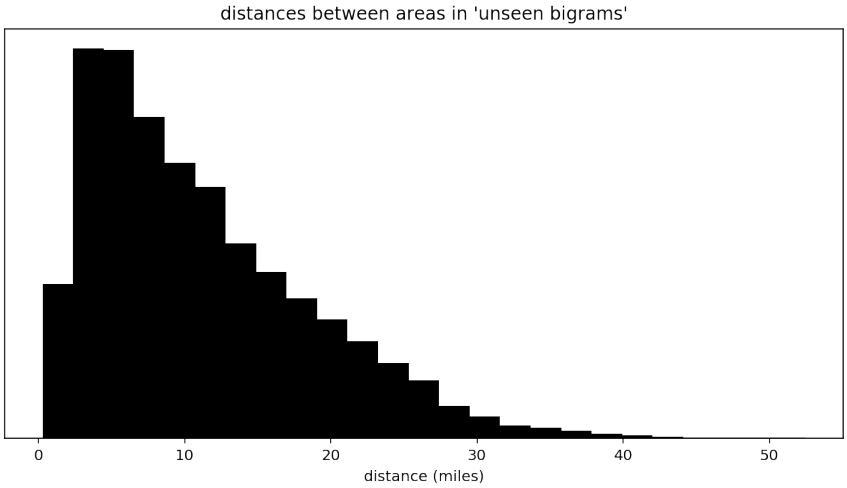


Figure 3-7: Distances between consecutive area pairs ("unseen bigrams") that occur in the synthetic stays trajectory dataset but that do not also occur in the real dataset. To evaluate the quality of the synthetic data, we verify that all distances between consecutive areas can be realistically traveled within the time period spanned by the trip that they represent.

The aggregate distribution of where users spend time is consistent with real data. The correlation between the distribution of aggregate time spent in each area, when comparing S and S' is ρ=0.936.

The distribution of the number of distinct places individuals visit within a time period is consistent with real data. The frequencies for the number of distinct areas in each stay trajectory in the real dataset, D, is shown in figure 3-9.

The numbers of distinct areas per stay trajectory are binned into 6 equal quantiles determined by the distribution of distinct areas per stay trajectory in D. Each bin is used as a category in the Pearson's chi-squared test for homogeneity, and expected frequencies are computed from the proportions of frequencies in D. We test the null hypothesis with a significance level of 0.05.

To test the methodology, we first test the real data sample S against D, resulting in a chi-square test statistic of 2.06 and p-value of 0.841 (see figure 3-10).

Testing S' results with a p-value of 0.429, allowing us to keep the null hypothesis that the distributions of frequencies are consistent between the synthetic and real data (see figure 3-11).

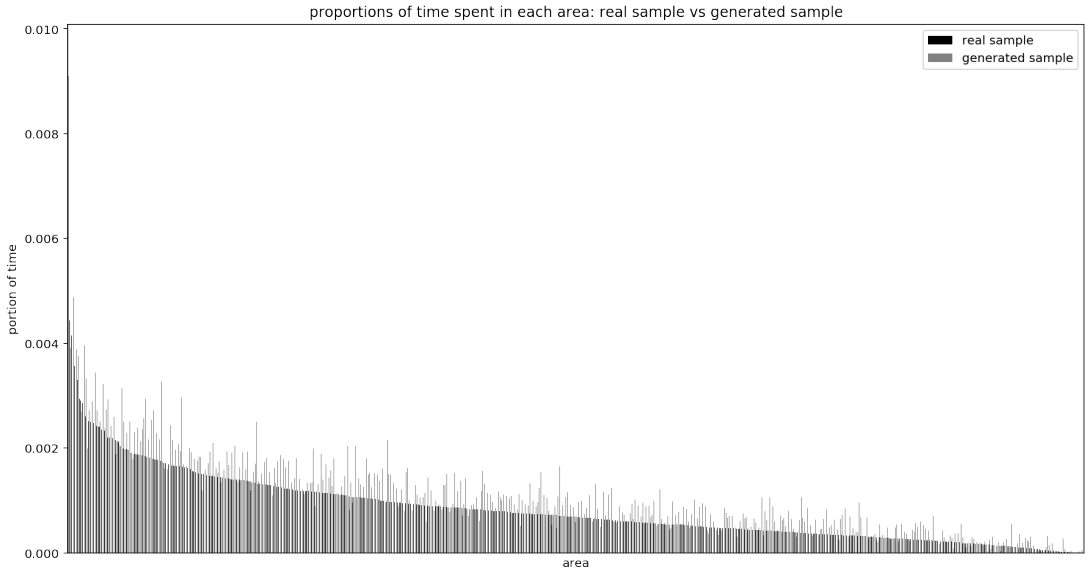


Figure 3-8: The proportion of aggregate time spent in each area for the real and synthetic datasets, S and S', respectively. The aggregate distributions of where synthetic users spend time should be consistent across the real and synthetic data. We compute the aggregate amount of time intervals users spend in each area for the real data, S, and synthetic data, S', and measure the correlation between their distributions. ρ=0.936.

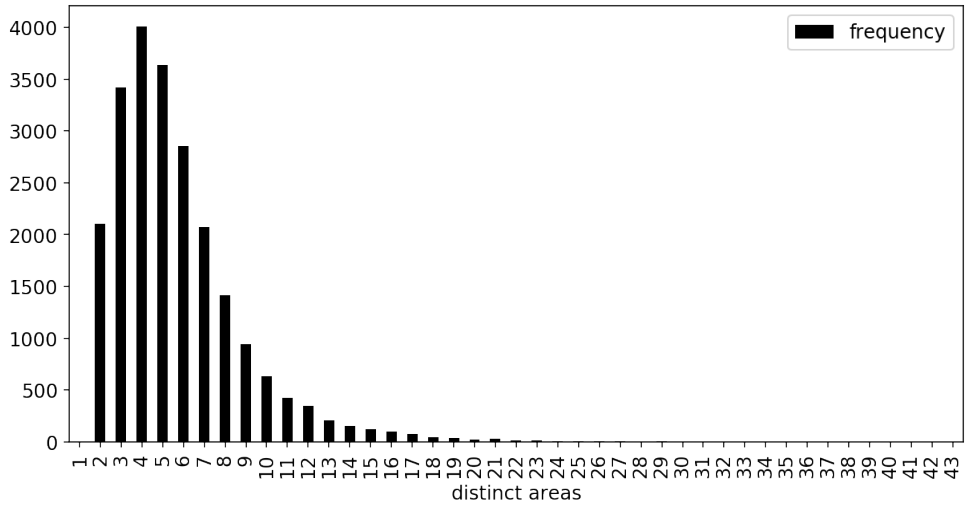


Figure 3-9: The distribution of the number of distinct areas in stay trajectories, counted over the entire real dataset, D.

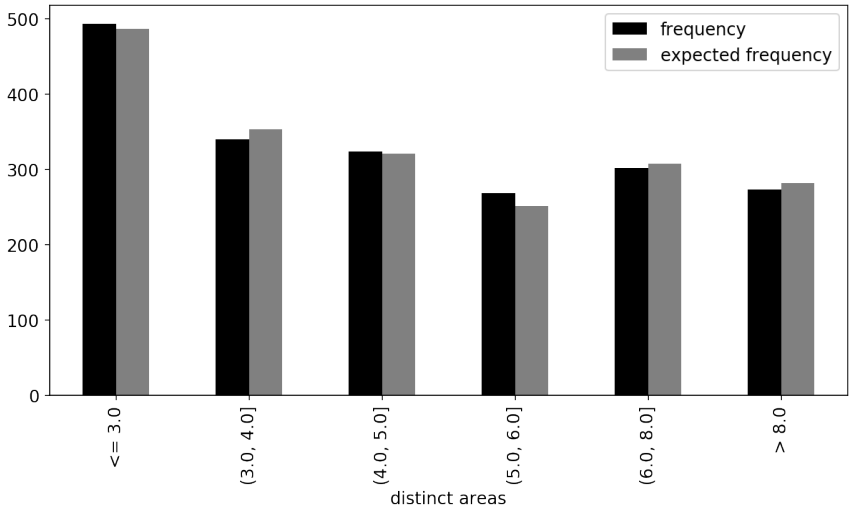


Figure 3-10: Comparing S against D with respect to frequencies of distinct places per stay trajectory. The chi-square test statistic is 2.06 and the p-value is 0.841.

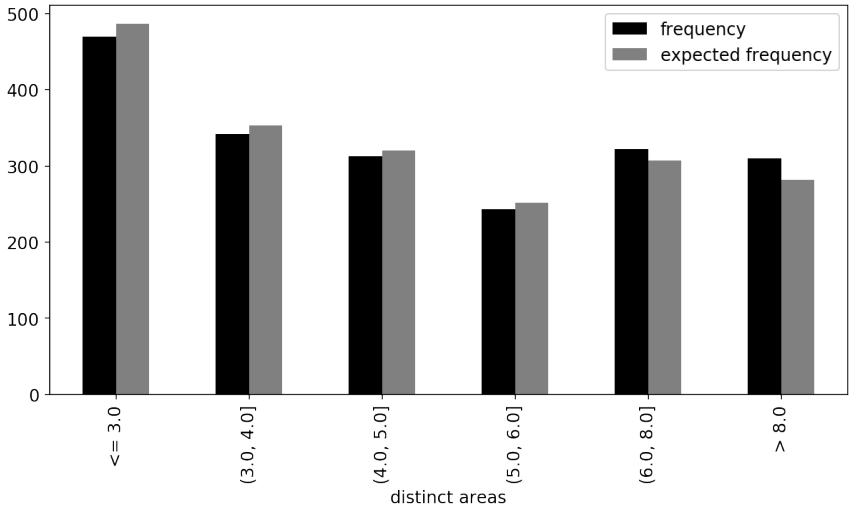


Figure 3-11: Comparing S' against D with respect to frequencies of distinct places per stay trajectory. The Pearson's chi-square test for homogeneity tests the null hypothesis that the distributions are consistent. The resulting p-value is 0.429, which is more than the significance level of 0.05, so we do not reject the null hypothesis.

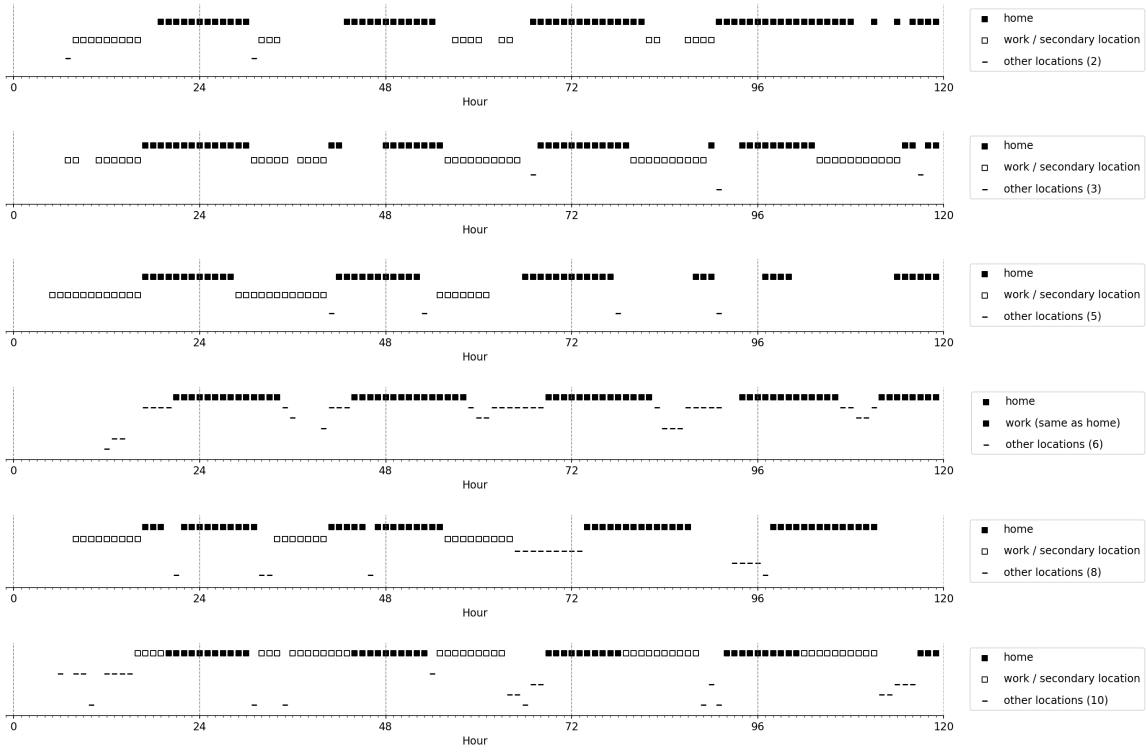


Figure 3-12: Individual mobility patterns plotted for a sample of synthetic stay trajectories. The plotted patterns can be compared to those for real stay trajectory data shown in figure 3-1. More plots for this model and others can be viewed in our open source evaluation code notebook: `https://github.com/aberke/lbs-data/blob/master/trajectory_synthesis/evaluation/evaluate_rnn.ipynb`

Mobility patterns at the individual level are retained. We show in figure 3-12 a sample of plotted synthetic stay trajectories in S'. For comparison, plots of mobility for individual stay trajectories from the real dataset, D, are shown in figure 3-1.

3.6.5 Privacy evaluation implementation and results

To evaluate how well the model meets our privacy criteria, we computed the minimum edit distance $m\text{-}dist(s', D)$ for each synthetic stay trajectory s' in the set S' to yield a corresponding set of minimum edit distance values.

$$\{s'_1, s'_2, ...\} \rightarrow \{m\text{-}dist(s'_1, D), m\text{-}dist(s'_2, D), ...\}$$

$Pr[m\text{-}dist(s', D) \leq m]$ is then estimated as the proportion of $m\text{-}dist(s', D)$ values such that $m\text{-}dist(s', D) \leq m$.

We similarly compute the $m\text{-}dist(s, D)$ values over the real sample, S.

$\{s_1, s_2, ...\} \rightarrow \{m\text{-}dist(s'_1, D), m\text{-}dist(s'_2, D), ...\}$

And estimate $Pr[m\text{-}dist(s, D) \leq m]$ as the proportion of $m\text{-}dist(s, D)$ values such that $m\text{-}dist(s, D) \leq m$.

We use these proportions to compare $Pr[m\text{-}dist(s', D) \leq m]$ to $Pr[m\text{-}dist(s, D) \leq m]$ for any possible value of m.

A caveat

The following evaluations compare values for $m\text{-}dist(s', D)$ and $m\text{-}dist(s, D)$. Since S is a subset of D and the edit distance between any s and itself is of course 0, computation for the distribution of $m\text{-}dist(s, D)$ values must avoid directly comparing any s to itself, in order to avoid yielding a distribution of 0's. Our implementation removes any s in S from D when computing $m\text{-}dist$ values for stay trajectories in the real sample, so that results for $m\text{-}dist(s, D)$ values are instead computed for $m\text{-}dist(s, D \setminus S)$. $m\text{-}dist(s'D)$ values for synthetic stay trajectories, s' in S', are still computed over all of D[14].

To make for a better comparison, the computations for $m\text{-}dist(s'_i, D)$ and $m\text{-}dist(s_i, D \setminus S)$ values skip any s' in S' and s in S with unique $\langle home, work \rangle$ label pairs. This is done because it is already effectively done for the computation over the real sample since any s in S with a unique $\langle home, work \rangle$ pair will not be compared to any other stay trajectory with that $\langle home, work \rangle$ pair when evaluating $m\text{-}dist(s, D \setminus S)$. We expect any stay trajectories with matching $\langle home, work \rangle$ pairs to be similar, so ignoring the unique $\langle home, work \rangle$ pairs for the real sample, S, without also doing so for the synthetic sample, S', would throw off what are intended as comparative distributions for $m\text{-}dist(s, D)$ and $m\text{-}dist(s', D)$[15]. The following evaluations still refer to the distributions of minimum edit distances as $m\text{-}dist(s, D)$ and $m\text{-}dist(s', D)$ for notational convenience, and we note that the resulting distributions still conveniently have the same number of total values. Later in this

[14]The s in S are removed from D only once, so any s with a duplicate in D will still be found in $D \setminus S$, resulting in a minimum edit distance of 0. This does occur in our dataset.

[15]Research from 2009 found that more than 5% of individuals in the U.S. working population have unique combinations of home and work census tract locations [104]. For this reason, excluding unique home and work location pairs from our privacy analysis would leave open to question the ability of the generated dataset to leak other sensitive location information.

section we describe how we account for these modifications with an additional evaluation.

Comparing probability distributions

At various levels of δ we find the value m for the real data, S, such that

$Pr[m\text{-}dist(s, D) \leq m] \leq \delta$

i.e. this is the value m such that the proportion of $m\text{-}dist(s, D)$ values less than or equal to m is δ .

We also find the corresponding value, m, for the synthetic data, S',

$Pr[m\text{-}dist(s', D) \leq m] \leq \delta$

And we compare these values of m directly.

The privacy criteria is satisfied when the m value for the synthetic data is equal or greater than the m value for the real data.

To evaluate the results over the full range of possible δ values, we used Q-Q plots. See figure 3-13 for the Q-Q plot evaluating S'.

The Q-Q plot matches the corresponding m values for the S and S' against each other, with values for S and S' on the x and y axes, respectively. Each point then shows data for a different δ, where the point's x-value is the corresponding m value for S, and the point's y-value is the corresponding m value for S'. A 45-degree line represents a matching distribution of values, and points on or above the 45-degree line represent where the privacy criteria is met. The values closer to the origin are the more important values, as these are for the smaller minimum edit distance values, m, where privacy risk is higher.

The values for our model closely track the 45-degree line, particularly at the smaller and more sensitive values. In other words, the distribution of minimum edit distance values for S' closely matches the distribution for S, as desired, without perfectly satisfying our privacy criteria.

We also established benchmarks by choosing values for δ as 0.01, 0.05, and 0.10 to find the

mean	31.398
std	13.176
minimum	0
1%	5
5%	11
10%	14
25%	22
50%	32
75%	40
maximum	75

Table 3.5: The distribution of $m\text{-}dist(s, D)$ values. Values for the 1%, 5%, 10% percentiles are used as benchmarks to evaluate S'.

corresponding m values for the real data, S. In other words, these values are cutoffs for the 1%, 5%, and 10% percentiles for the distribution of $m\text{-}dist(s, D)$ values. These benchmark values are 5, 11, and 14, respectively, and are shown in table 3.5. The corresponding values for S' are 5, 10, and 14. Values for other models are shown in table 3.4.

Additional evaluation

In order to handle the caveat described above, where stay trajectories with unique $\langle home, work \rangle$ pairs are ignored, we do the following. We used $\mathbf{M}$ to generate an additional synthetic dataset, S'', with the same distribution of $\langle home, work \rangle$ pairs as S' and S. We then compared the distributions of minimum edit distances between S' and S'', using all stay trajectories in the evaluation, including those with unique $\langle home, work \rangle$ pairs. In other words, we computed $m\text{-}dist(s'', S')$ for each s'' in S'' over all s in S'.

We then evaluated modified privacy criteria, $Pr[m\text{-}dist(s'', S') \leq m] \leq Pr[m\text{-}dist(s, D) \leq m]$, using the benchmarks already established for $Pr[m\text{-}dist(s, D) \leq m]$. The 1%, 5%, and 10% percentile cutoff values are 8, 15, and 20, respectively, satisfying the modified privacy criteria.

3.7 Discussion and Conclusion

In this chapter I presented a model to use real and private location data that was sampled from the population in order to generate synthetic data that is representative of the real

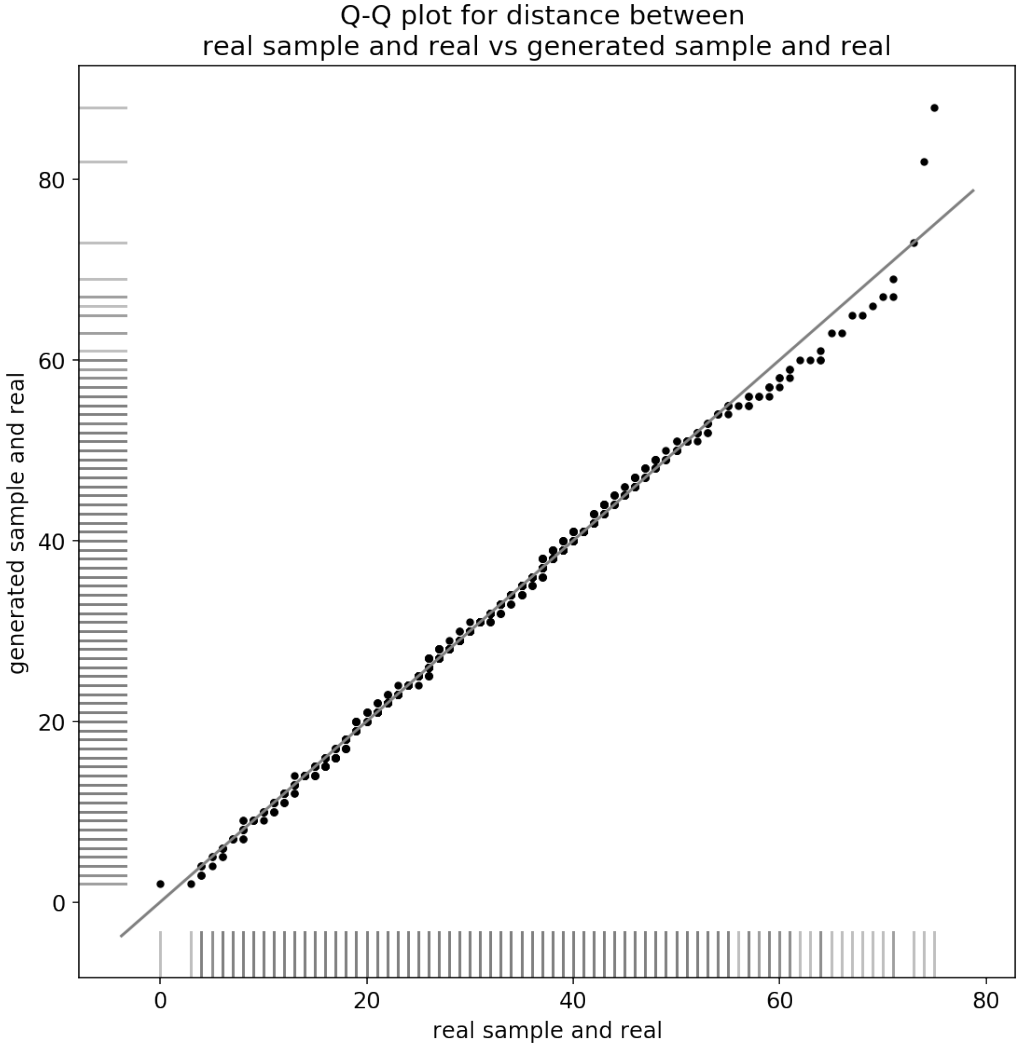


Figure 3-13: A Q-Q plot comparing the distributions of $m\text{-}dist(s, D \setminus S)$ values and $m\text{-}dist(s', D)$ values. The $m\text{-}dist(s, D)$ values are on the x-axis and the $m\text{-}dist(s', D)$ values are on the y-axis. To create the points in the plot, the values of the sets $\{m\text{-}dist(s_1, D), m\text{-}dist(s_2, D), ...\}$ and $\{m\text{-}dist(s'_1, D), m\text{-}dist(s'_2, D), ...\}$ were sorted and matched by their ordered index. Each point on or above the 45-degree line represents where $Pr[m\text{-}dist(s', D) \leq m] \leq Pr[m\text{-}dist(s, D) \leq m]$ for the corresponding value m, satisfying the privacy criterion.

data and preserves its utility, while also preserving user privacy. The model was designed to generate data for synthetic users with a desired distribution of home and work locations, in order to serve applications that model the entire population.

We intend for this work to be extended to allow for privacy-safe public usage and democratization of the information provided by location data. Datasets such as the LBS dataset used in this work are often kept private to serve the business interests of the companies that collect them or because their publication would risk user privacy. Their utility then cannot be fully realized, as they cannot be used by the many organizations and researchers whose work could benefit from them, and whose work could benefit the greater good. Yet synthetic datasets could be published as a public good. This need not detract from the bottom line of private companies who collect the real data from which the synthetic data is produced. While synthetic datasets that preserve the properties of real datasets can be useful, private companies could still profit from exclusive access to the real data which provides precision and accuracy and serves specific use cases that synthetic data cannot.

The presented work applied our methodology to a small and sparse dataset, which necessitated limited levels of spatial and temporal precision for the generated synthetic output. However with larger datasets, the methodology we presented can be applied for higher levels of precision.

Moreover our approach to synthetic data may offer a unique opportunity to safely use larger datasets and combine multiple datasets from various sources. Data from different sources might contain different or overlapping data points for the same users, collected at different times and locations by various applications or devices or methods. In most location data use cases this would incite the need to merge data sources so that each user is only represented once in the dataset [105]. Merging data for users in this way is difficult to do with accuracy. Moreover, successfully doing so further risks the privacy of users by adding more information and sources of uniqueness to their trajectory data. However in our use case, where data is used to train a model to generate synthetic data, multiple datasets with possibly duplicated data could be combined and leveraged to improve the model without a need to merge them.

The methods presented in this chapter can also be applied to other types of location data. Our preprocessing transformed location based services data, with real-valued latitude and

longitude points, into sequences of discrete tokens, which we called "stay trajectories". Yet many other types of mobility data already represent visits to a discrete set of places and are then already in a similar form. Common examples include datasets for station visits within a transportation system and check-ins at points of interest ("POI" data).

Future works can extend our utility and privacy framework to meet the needs of these other location data types and sources. The framework should also be extended to address the additional levels of precision needed if this work is applied to larger and higher resolution data. Extensions might include making the metrics more granular in terms of inspecting where users spend time and inspecting temporal and spatial attributes together, as well as adding additional metrics informed by the cited related works.

Furthermore, this work addresses a lack of robust criteria and methods to evaluate privacy for synthetic location datasets. This chapter presented new "indistinguishability" criteria to evaluate the privacy preserving properties of synthetic location data and fill this void. The criteria are generalizable and future work can expand upon them.

Part II

USING LOCATION DATA TO ADDRESS THE COVID-19 CRISIS

In the midst of this thesis work, the COVID-19 health crisis became a global pandemic. At the same time, location data emerged as an important tool to better understand and help mitigate the crisis. As a researcher working with this data, I redirected my work towards these efforts.

This part of the work begins with privacy risks and challenges when using location data for contact tracing. It also shows how location data can be used to address the health crisis while better preserving privacy.

Overall this work demonstrates how location data can serve the public as an important and useful tool, particularly when contending with a public health crisis.

Chapter 4

Contact tracing technologies: Methods and trade-offs

In the spring of 2020, governments around the world considered the deployment of contact tracing technologies to help contain the spread of COVID-19 and mitigate its economic impacts. Combined with increased testing, effective contact tracing offered the opportunity to improve policy decisions by providing information to help safely re-open economies and intervene only upon the detection of new outbreaks.

However, it was not yet known whether contact tracing technologies could deliver their desired outcomes. They would need to be widely adopted and accurate in order to be effective, and they would need to provide enough information about their users to health authorities or governments in order to guide future policy decisions. These challenges raised both technical issues and societal issues, as deploying effective contact tracing technologies risked jeopardizing individual privacy rights and freedoms. **The use of location data was central to many of these issues.**

This chapter presents work done in the spring of 2020, as governments and societies grappled with these challenges, and is adapted from a whitepaper I coauthored with Kent Larson [106]. This chapter describes various ways contact tracing technologies can be designed, and how each design decision leads to different trade-offs between their potential accuracy, adoption, usefulness, and privacy risks. We did this work in order to inform decision makers who

sought these technologies and the communities who might then use them. We believed it important for the public to understand how contact tracing systems use location data, and their alternatives, because their widespread adoption could drastically impact privacy and change how people move in public.

This chapter begins by providing an overview of contact tracing technologies already developed. It explains how they work and then how alternatives could work instead. The following sections cover differences in how location data is sourced and used to detect contacts, whether the flow of information is centralized or decentralized, how COVID-positive cases are reported, how exposure risk is assessed and how the system's users are impacted. These sections also cover how these differences lead to trade-offs between accuracy, adoption, usefulness, and privacy.

The goal of this work was to help inform readers about the trade-offs of contact tracing technologies as well as raise critical questions. Can any of these technologies be useful enough to be worth their trade-offs?

4.1 Background: Contact tracing & technology

Contact tracing is a longstanding public health strategy for reducing the spread of infectious disease by identifying people who may have been exposed. Traditionally, contact tracing involves asking infected people to disclose where they have recently been and with whom they may have come in contact, and then following up with those contacts. But this process is labor-intensive and many people cannot sufficiently recall all of their recent movements, or know all of the people they came in contact with.

Technology can help make this process both more efficient and more accurate, and help scale up existing human-driven contact tracing initiatives. For example, location data collected by mobile phones can aid a patient's memory and increase the accuracy and speed of traditional contact tracing interviews. Mobile applications can also be used to connect people with resources for getting tested or provide guidance for quarantine or other means to reduce the spread of infection. This overview focuses on systems that go beyond assisting human-driven contact tracing interviews. It focuses on technologies that use location data to automatically

identify and notify individuals who may have been exposed and assess risk.

4.2 Early efforts for contact tracing

By early spring of 2020 there were already significant efforts to use technology to scale up the process of contact tracing in order to control the outbreak of COVID-19, including projects coordinated by governments, open source communities, and private companies. Notable examples had already been deployed by governments in Asia.

South Korea effectively traced travel routes and contacts for infected patients by using a range of data sources, such as in-person interviews, GPS data from cell phones, and credit card transactions [107, 108], to all of which they have legal access. There were multiple websites and smartphone apps that published this location data with timelines and maps, including granular details such as which bus someone took, when and where they got on and off, or whether they were wearing a face mask [109, 110]. The government also broadcasted emergency alert information to nearby citizens whenever new cases were discovered in their districts. This provided a basic way to inform people of their risk of exposure, based on whether they may have crossed paths with those infected.

In China, the "Alipay Health Code" system involved a mobile app that created colored QR codes for each user, where a color of red, yellow or green indicated the user's exposure risk level determined by the system, and dictated their quarantine [111]. Exactly how exposure risk was determined was not public, but involved combining people's personal information with their recent travel histories and locations. Using this app, or systems like it, became a de facto requirement in hundreds of cities across China, where scanning a green QR code was required to enter many buildings, or travel, or return to work.

In Singapore, The Ministry of Health and Government Technology Agency launched the TraceTogether mobile app for more targeted contact tracing [112]. Instead of using geographic location data to detect whether people were in the same place at the same time, the system was designed to use Bluetooth signals to detect whether two app users came into proximity of one another. The app was voluntary and its limited use raised the question of whether opt-in systems could be effective or whether there would need to be ways to

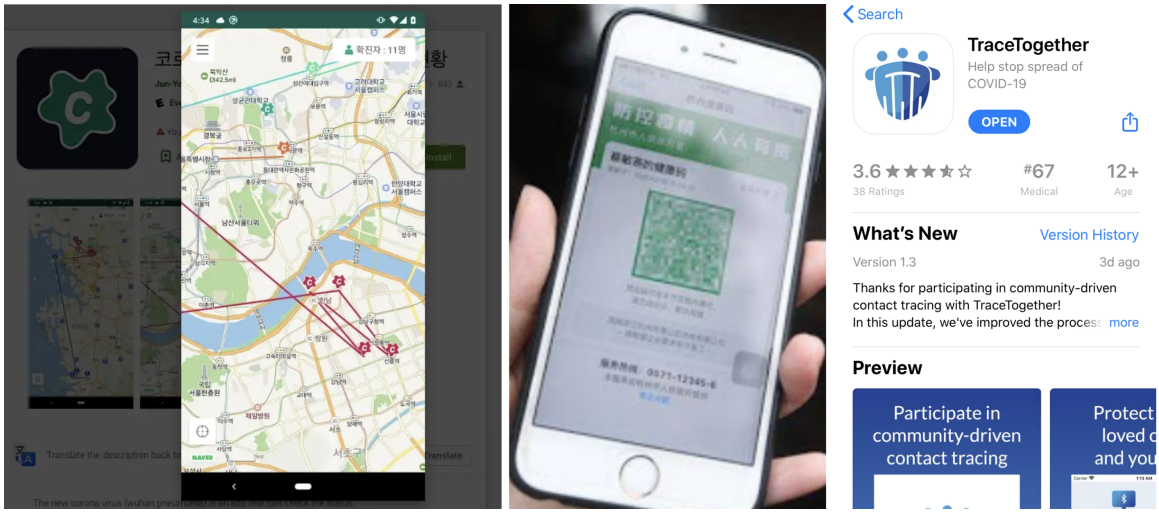


Figure 4-1: (Left) In South Korea, multiple apps and websites published maps and detailed timelines for infected patients' travel histories. (Center) The AliPay Health Code app used in China. (Right) Singapore's TraceTogether app.

incentivize their use.

In all of these described systems, the collection and management of their users' data was centralized, allowing the governing authority to more effectively act upon it.

4.2.1 Independent projects and research

Many more independent projects and proposals were developed using data in similar ways to those systems implemented by governments, but with more privacy-preserving and decentralized technology designs. Projects such as CoEpi [113], Covid-Watch [114], DP-3T [115], and PACT (Private Automated Contact Tracing) from MIT [116] were designed to limit centralized collection of people's private data in order to limit its potential abuse. Many of these systems focused on the use of Bluetooth technology because it provides ways of precisely detecting whether users of an app come into contact with each other without exposing other sensitive information that location histories can reveal.

Shared protocols

These projects were designed to interoperate with shared protocols (e.g. TCNCoalition [117]), so that information about infection status could be shared across users of different mobile applications and systems. How a user's infection exposure risk level is assessed and shared could then be tailored to each project's specific objectives.

From independent projects and research to adoption

The ideas from these independent projects and proposals were incorporated into a framework created in a joint effort by Apple and Google. The Apple-Google framework provides a software layer that interfaces with Bluetooth, on top of which software developers who work on behalf of various public health authorities can build apps. The framework is designed around providing security and privacy for users, and the early drafts for the Bluetooth and cryptography specifications [118] closely followed the suggestions of the privacy-preserving research proposals, such as PACT and DP-3T. (The appendix section B.1 provides a high-level description of how these specifications work.) Their intentions also seemed consistent with the desire to protect users' data from the centralized collection by governments. This initially put them at odds with French and British health authorities, which had plans for more centralized contact tracing systems [119, 120].

Using Bluetooth to create effective contact tracing systems had previously been difficult due to compatibility issues between Google Android and Apple iOS devices, as well as iOS limitations on the continuous broadcasting of Bluetooth signals (due to privacy considerations). Needless to say, the new Apple-Google framework changed this by providing an interface to more easily use Bluetooth for contact tracing, and also improving interoperability between Android and iOS devices.

4.3 Implementation differences and trade-offs

The previously described contact tracing technologies and their possible alternatives differ in a number of ways including:

- How data is used to detect contacts
- How trust and the flow of information is managed
- How positive cases are reported
- How exposure risk is assessed and how it impacts users

These methodological differences in turn lead to trade-offs between:

- Adoption
- Accuracy
- Usefulness to public health authorities and decision makers
- Usefulness to individuals
- Privacy

The following sections discuss these trade-offs. But first, let's consider privacy in terms of *whose* privacy is protected and *from whom* privacy is being protected. We can consider two categories of users for contact tracing apps: users who report as infected and share their data, and users with whom they may have come in contact with and may therefore have been exposed. We can consider three different notions of privacy for users: (1) privacy from authorities administering the system or app, (2) privacy from potential contacts, and (3) privacy from anyone else. "Anyone else" might include snoopers trying to collect information about individuals, or it might include companies increasing their existing collection of user data to help them better target ads or for other means of private profit.

All of the presented contact tracing projects and research proposals assume that users with a positive COVID-19 test result must surrender some privacy when reporting their data. However, the amount of privacy they surrender, and to whom, varies depending on implementation.

To further protect user privacy, any contact tracing system should only collect data relevant to the disease or situation it is targeting and data should be deleted after a predefined period

that experts consider medically relevant. These data minimization and storage limitation measures are not specific to any of the alternative technology designs or methods later discussed, they are simply standard good practice as well as part of GDPR compliance [121].

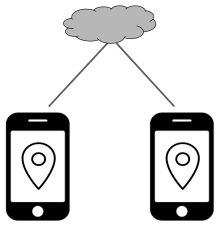
4.3.1 How data is used to detect contacts: Location data versus Bluetooth co-locations

Different forms of data can be used for contact tracing including location data such as timestamped geolocation coordinates, or data collected via Bluetooth signals. There are also different ways this data can serve contact tracing, for example, by creating maps and aggregate statistics, or detecting whether two people have come into contact. In what follows I explain these different data sources and ways to use them, as well as their trade-offs.

How data is collected and what makes it useful

People's location histories are commonly recorded by applications installed on their mobile devices in the form of timestamped geolocation coordinates, often collected via GPS. Co-location data collected via Bluetooth is different. Devices that broadcast and receive messages over short-range Bluetooth signals can exchange data peer-to-peer when they come in close enough proximity to one another. This data exchange provides information about whether people were co-located rather than just their geographic locations. This section describes how this can be useful for privacy-preserving contact tracing.

Once collected, geolocation and Bluetooth data can be used to scale up contact tracing efforts in multiple ways. One way that geolocation data can be used, but that Bluetooth data cannot, is to create maps and timelines or aggregate statistics about when and where people went before they were diagnosed as infected. These data and the resulting visual-



izations can be useful to public health agencies, and making this information public can help inform other people of their exposure risk. This approach was used in South Korea, with the release of detailed timelines showing the locations of infected individuals. However, publishing this detailed information risks exposing private information about the infected people reported on, and risks the stigmatization of the businesses or communities that these people visited. This data could instead be more safely anonymized and aggregated, but there is a trade-off: the data is more informative when it is more detailed, but safer for use from a privacy perspective when it is more aggregated and less detailed. Even in aggregated form, however, geolocation data can be used to create heatmaps and for statistical analysis to better understand geographic transmission flow and trends of disease outbreaks.

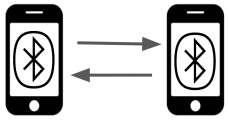
Another use case for both geolocation data and Bluetooth co-location data is in more targeted person-to-person contact tracing.

Trade-offs: Accuracy, usefulness, and privacy

Location data approach: Geolocation data can be used to estimate an individual's disease exposure risk by detecting whether their mobile device reported a location near an infected person's at roughly the same time, and noting how long they were in that place together.

One drawback to using geolocation data in this way is that it often relies on GPS which suffers from limited accuracy in dense urban areas or indoors, and cannot pinpoint which room or floor in a building an individual was, making it less useful for detecting if people came in contact. However, GPS accuracy can be somewhat improved when combined with data logged by wifi routers.

Another drawback to using geolocation data is privacy, as location histories can reveal private



and sensitive information about people. This is the case even when data is anonymized, because statistical methods can be used to reconstruct location histories and re-identify people [32]. Redaction can help mitigate this risk. For example, systems like SafePaths [122] allow health providers or users to retroactively redact their location histories before sharing them. Apps could also allow users to proactively set places and times when their data will not be recorded at all.

Bluetooth co-location data approach: Some of the issues with location data can be resolved by using Bluetooth co-location data for the more targeted person-to-person contact tracing. (This approach is used by TraceTogether, CoEpi, COVID-Watch, PACT, DP-3T, the Apple-Google framework, and others.) Applications installed on users' mobile devices use the Bluetooth Low Energy (BLE) protocol to broadcast IDs and listen for IDs broadcast from other devices. Each app records information about their broadcast IDs and received IDs. Since Bluetooth signals are short-range, the apps can only exchange IDs when devices come in close proximity of one another, serving as a good proxy for whether users came in close enough contact to transmit disease. Users who later report a positive infection status can share information about the IDs their app broadcast or received (depending on the implementation). Exposure risk for other users can then be assessed by whether their app exchanged IDs with infected users' apps.

One of the benefits of using Bluetooth for more targeted contact tracing is that it can allow a system to better preserve user privacy. While location data can expose people's private information, using Bluetooth mitigates privacy risks by detecting when people come into contact without using location data. The app that broadcasts IDs from users' devices can generate the IDs in such a way as to make them look random, and it can change IDs frequently. It is still possible to track people between the places they go by recording the IDs their devices broadcast and the locations where they are broadcast. However, this requires devices to be located at the places people go in order to receive the broadcasts, and the places people are tracked would be limited to those locations. Tracking people in this way

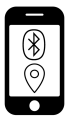
is further made difficult when the broadcast IDs change.

Another benefit of the Bluetooth approach is that it can overcome the accuracy issues of GPS. The Bluetooth signal is short-range and degrades when crossing between the walls and floors of a building, enabling a more precise detection of whether two people were in a shared space. A measure of signal strength can also be used as a proxy for how closely two people came into contact and this measure can be used to better assess exposure risk.

Yet by itself Bluetooth data may not be accurate or useful enough, as its lack of location information also means a lack of contextual information. For example, Bluetooth data alone cannot differentiate between contact with an infected user while in a closed setting like a restaurant, where many people may touch surfaces, versus outdoors. These very different settings imply different levels of exposure risk.

Another trade-off between privacy and usefulness to consider for Bluetooth-based systems is that Bluetooth can only detect when people were in the same place at the same time. It may miss situations when people shared common spaces at slightly different times. In these cases disease may transmit across commonly touched surfaces (fomites), such as grocery check-out counters. Geolocation-based approaches can be more accurate in this regard because they can account for time ranges when comparing time and location to detect points of contact. To resolve this issue for apps that use Bluetooth, dedicated beacons that act as signal repeaters can be installed at common locations. These beacons could repeat the signals broadcast by app users that came near them for a limited time period, so that the next app user that comes near the beacon also receives the signal. However, associating Bluetooth beacons with dedicated locations, and having these beacons listen to users' broadcast signals then degrades the central privacy feature for using Bluetooth: signals received about co-location are not associated with locations. If the beacons store information about the signals received, this information could be used to determine where someone was, and who else was there with them.

The potential use of beacons also raises new privacy questions. Bluetooth beacons are already used by retailers in stores to track customers' behaviors in order to better sell their products [123]. We can imagine a future where beacons like these are as common in our environments as the ubiquitous collection of GPS location data from our mobile



devices. Building contact tracing technologies that cause our devices to constantly broadcast Bluetooth signals may bring about this future more quickly. That is, even though Bluetooth-based systems may better preserve location privacy right now, building these systems may create more precise ways to track people in the future.

Hybrid approach: Using Bluetooth co-location data in combination with geolocation histories would likely result in the most accurate and useful contact tracing systems. Bluetooth co-location data can be used for the more precise detection of contacts, while GPS location histories can provide data for aggregated statistics and heatmaps.

Location data can also improve contact detection done via Bluetooth. For example, when an app using Bluetooth exchanges IDs with another app, it might also record metadata, such as the time and location where those IDs were broadcast or received. If those IDs are later shared by an infected person in order to indicate exposure risk to their contacts, a system can then connect those IDs to the time and place an app stored them. This can provide useful context about where a user was at these points of contact in order to better assess exposure risk.

However, while this exchange may make an app more useful, it also re-introduces the privacy issues associated with location data, as co-locations are then reconnected to locations. An app could mitigate this risk for its user by only storing locations locally and never sharing them, so that only the app's user would see where it came into contact with infected people. However, this does little to preserve the privacy for the infected users who shared their data, as their locations will then be shared with their contacts who could then identify them.

To further improve the accuracy of GPS location and Bluetooth data, new data sources were also proposed for contact tracing [124]. For example, high pitched audio signals above the human audible range could be transmitted between contact tracing apps instead of Bluetooth signals. These audio signals are more likely than Bluetooth signals (transmitted over radio waves) to be stopped by walls, ceilings and floors, and can be used to more

accurately measure distance between devices because they travel through air more slowly than radio. More sensors that measure air quality, such as the barometers that measure air pressure in some smartphones, can be used to provide more information on whether people were in a shared space together, sharing the same air [125]. Combining a variety of data sources may be key to reducing the rate of false negatives and false positives and increasing the overall accuracy and usefulness of any contact tracing system.

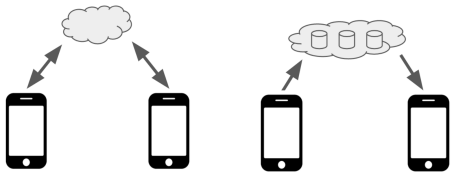
In addition to data directly collected by smartphones, data sources such as credit card transactions, transit pass records, or CCTV footage (all of which have been used in South Korea's contact tracing efforts), also provide useful information to improve accuracy. However, each of these data sources also present trade-offs between the added accuracy and usefulness they provide, and privacy.

Trade-offs: Adoption

An important issue to consider for Bluetooth-based systems is adoption. Systems that rely on using Bluetooth to detect contacts will require the mass adoption of a new mobile app before they can be useful, while this is not necessarily the case for systems using geolocations. Users of Bluetooth-based contact tracing apps need to exchange enough data via the apps, before infected users report their data, in order for their contacts to be detected. In addition, a substantially large portion of the population needs to consistently use the system in order for it to provide enough useful information to the people who do use it. Even in Singapore where there is a government app (TraceTogether), fewer than 20% of people had downloaded it months after it was released. If only 20% of people use an app, the system can only hope to detect about 4% (0.2 x 0.2) of the encounters between people. Needless to say, too many points of contact with infected people will go undetected for an app like TraceTogether to have a meaningful impact.

On the other hand, geolocation data is already collected from mobile devices by a variety of apps and companies, and two users do not need their devices to directly interact via an app in order for this data to be useful for contact tracing. This data can even be used before the adoption of new apps. For example, users can export their Google location histories[1], or

[1] `https://takeout.google.com/settings/takeout`



companies might share or be compelled to share, the location data they have been amassing.

Countries such as China and Israel made use of data already collected from people's devices, rather than allowing users to opt-in to their surveillance. This data provided a way to scale contact tracing efforts as well as enforce quarantines by monitoring whether people stayed at home. Using location data in this way can make a system useful with the immediacy needed to effectively stem the rate of further infections, but may forfeit the privacy and rights of the citizens who did not explicitly consent to being tracked by the system.

4.3.2 How trust and the flow of information is managed: Centralized versus decentralized

The contact tracing technology systems used by South Korea, China (AliPay Health Code), and Singapore (TraceTogether) are centralized, meaning a single entity collects location, co-location or other data from all users whether or not they have positively tested as infected. These entities also control the flow and use of this information. For example, China's system could use location histories from all of its users to find similarities and determine which users were more likely exposed to infected users.

Similarly, Singapore's TraceTogether app, which uses Bluetooth to exchange IDs, keeps a database linking IDs that users broadcast to users' identities and phone numbers. When users are diagnosed as infected, they must then share the IDs that their app received from other users (their likely contacts) with TraceTogether's central server. These IDs are then connected back to the information stored about these users so that authorities can learn who was exposed and reach out to them via their phone numbers.

Decentralized systems, such as those proposed by CoEpi, Covid-Watch, PACT, and enabled by the Apple-Google framework, work slightly differently. When users are diagnosed as infected, they (optionally) share their data. This data may even be shared to a central

database. What makes the system decentralized is that other users can then download or query this data without sharing their own information. The data they receive can then be used to assess their exposure risk within their app.

Trade-offs: Privacy and usefulness

The centralized and decentralized system designs differ in terms of whose privacy is preserved and from whom. In each case, infected users give up some privacy when reporting their data, but with a centralized system design, they need only share this data with the authority managing the system. Centralized designs can work well for users if they trust the authority managing the system because by collecting data from all users, the system can do the work of finding points of contact while protecting users' privacy from others. But no users have privacy from the central authority, which may be a government, an organization or a company. The authority may reserve the right to act on its knowledge of contacts, not only to notify people of their exposure risk, but possibly to ensure that exposed contacts quarantine or limit their travel, as happened in China. The amount of information and level of control afforded to governments may make the centralized system most useful for them and their citizens and therefore most desirable. Or it may raise concerns. In countries like the US, citizen concerns about forfeiting privacy and control to a central authority may stymie the adoption of a centralized system, making it less effective.

In the case of a decentralized system, users can query the system to find whether they had contact with infected users without sharing their own information. This makes it difficult for an authority to gain an overall view of which or how many users came into contact with infected users. A decentralized approach can increase privacy for most users, but at the potential cost of privacy for infected users who share their data and whose data is then accessible. Some systems use additional privacy protection measures, such as mixnets[2] and private set intersection protocols[3], to limit the amount of information anyone can have about infected users' data, as well as the amount of information users can expose by querying the system [126, 127]. However, these additional privacy measures add implementation complexity.

[2] `https://en.wikipedia.org/wiki/Mix_network`
[3] `https://en.wikipedia.org/wiki/Private_set_intersection`

We might also consider systems that notify contacts of contacts. Users will often become contagious and expose their contacts to risk well before they report as infected and notify their contacts. During this time, their contacts may expose other users, who could later be notified of their risk as well. Centralized systems may be better equipped to handle this because they can detect contacts of contacts, assess their risk, and notify them, without relying on the users who were the directly exposed contacts to take any extra action or share new information.

In general, the decentralized approach provides users with more privacy and autonomy, and authorities with less information and control.

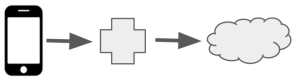
4.3.3 How positive cases are reported

Many systems were designed to use information only from positive test results submitted by trusted health agencies to inform exposure risk. This submission process can be distributed in a secure way. For example, Covid-Watch and PACT proposed the idea of "permission numbers". With this scheme, each testing authority generates a list of permission numbers that are distributed to health providers that are authorized to diagnose individuals. Each permission number is "use once", meaning it is used to authorize the upload of information from one diagnosed individual. And permission numbers are generated in a way to make them nearly impossible to guess, keeping the system secure from unauthorized data uploads.

Other system designs allow users to self-report symptoms. For example, the CoEpi team developed an app for users to voluntarily report symptoms and symptom severity from the app. This can allow for more immediate identification of positive cases and exposed contacts, and may be ideal for areas with limited testing resources, but this feature also has trade-offs.

4.3.4 Trade-offs: Usefulness

Including self-reported data can enable any system to more quickly scale its collection of data without the bottlenecks of hospital visits and limited access to certified test results. This could allow a system to be more useful, more quickly, to more users. On the other



hand, including self-reported data could degrade the quality and integrity of the system's data, as people may misdiagnose their own illness or share low quality or false data, either unintentionally or intentionally. This could decrease the accuracy of the system as well as users' trust in the system, making it less useful.

Systems that use self-reported data can interoperate with systems that only use data shared by authorized healthcare providers, which creates a middle ground. For example, Covid-Watch and CoEpi both planned to implement the same protocol to allow data sharing across their users but CoEpi uses self-reported data while Covid-Watch does not. Metadata can be connected to reported data points to indicate whether these data points came from a self-report or a trusted health provider. Different applications may then choose to treat this data differently. For example, an app could ignore self-reported data or use it as a weaker indicator than data submitted by health providers in its assessment of exposure risk. A system that successfully leverages self-reported data in combination with official test results could prove most useful.

4.3.5 How exposure risk is assessed and how it impacts users

Contact tracing systems can differ in how they assess exposure risk and present this information in apps for their users. These systems can also use risk assessments to limit the mobility of their users. For example, China's AliPay Health Code app uses color codes to show users their assessed risk levels. The colors and associated QR codes were used to limit and further track the mobility of the app's users. This feature was very useful to the Chinese government for managing the health crisis, but it limited the freedom and privacy of citizens. Another issue was transparency. Users were not told how their risk level was determined, so the color code was not very informative. And because they cannot control or contest the color codes fairness was an issue as well.

Other apps could tell a user the estimated number of times they came in contact with infected people, or their estimated total contact duration, and evaluate exposure risk based

on these numbers in a more transparent way. Apps might even show a user when or where contact with infected people occurred. As previously described, this can make an app more informative for users, but it can also present privacy issues by potentially exposing the identities of infected people to their contacts.

Other systems may incorporate more personalized information and AI into their risk calculation. For example, the MILA group[4] designed a contact tracing system for the Canadian government with a machine learning component to predict a person's probability of having COVID-19 based on their medical information in combination with location data. This was done to more intelligently estimate personalized risk levels for users.

All contact tracing systems will have limits to their accuracy due to the limitations of technology and the complexity of human interactions. Potential false positives or false negatives require careful handling. Reporting false positives can be harmful for users who might then go to a hospital to seek a test, or who are wrongfully directed to quarantine. Similarly, false negatives are also an issue. These can occur when points of contact are missed, for example, because someone does not consistently carry a mobile device or use an app, or because the system is not sensitive enough. If apps give users a false sense of security when false negatives occur, users may then expose themselves or others to risk.

There is then a trade-off between providing users with sufficiently detailed information to demonstrate a level of confidence needed to make recommendations for testing or quarantine, versus providing less precise indicators of exposure risk to hedge against wrongfully reporting information.

4.4 Risks and questions beyond contact tracing

For any contact tracing technologies, we have to question how useful they can really be, and how to even measure their effectiveness. We also have to question what their deployments

[4] `https://mila.quebec/en/covid-19/`

will mean for privacy and freedom in both the immediate and distant futures. Even research proposals that use Bluetooth protocols and decentralized designs to best protect users' privacy from central authorities may still create new ways for people to be tracked. Can the potential benefits of contact tracing technologies be worth their trade-offs?

Even if the technology for these contact tracing systems could work with a high degree of accuracy, we must question whether they could provide a solution to the COVID-19 epidemic. Suppose there are far more asymptomatic cases than confirmed cases; is the tracing of only those who test positive even useful? The most accurate systems would likely require adoption of a new app to use Bluetooth. Researchers estimated that a majority of the population would need to use a Bluetooth app for it to be useful [128]. Yet only about 1 in 5 people in Singapore used their government's TraceTogether app in the midst of the epidemic. Can we expect enough people to opt in to such a system, or will governments need to enforce or otherwise incentivize its use?

Moreover, these technologies can only be useful if the people they notify about potential exposure risk are able to be tested, get treatment, or self-isolate. Could these options be made available and affordable for enough of the population?

This chapter presented alternative methods and trade-offs to consider when building contact tracing technologies; but ultimately, how these systems are built and used rests on the consideration of societal questions about privacy and freedom and access to health services. If we do not think about these questions as a society and intentionally design our technologies and policies to address them, then these decisions might be made for us. Consider how the QR codes from China's AliPay Health Code were used. We might imagine a future where presenting an app that shows a low exposure risk or confirmation of good health becomes necessary to board a train or airplane, or enter a building or place of work. Then, even systems that were designed as opt-in could become effectively required.

In the spring of 2020, contact tracing technologies were already being built and used to address the health crisis in ways that risked individual privacy and freedom. When weighing the trade-offs for technologies such as these, we must also consider that the risks they pose can last beyond a time of crisis.

Chapter 5

Assessing disease exposure risk with location data: A proposal for cryptographic preservation of privacy

5.1 Introduction

In the previous chapter I outlined alternative contact tracing technologies and their trade-offs. In the months preceding that work, this range of alternatives had not yet been fully developed or explored. This chapter summarizes work done during those earlier stages of the COVID-19 crisis, as my colleagues and I took on the task of exploring and developing alternatives; this work was also described in a paper I coauthored with colleagues[1] [127]. During that time, it became clear that contact tracing could serve as a crucially effective tool as health entities, communities and governments attempted to contain the viral outbreak. It was also clear that location data collected from personal devices could enable contact tracing processes to scale.

There were already digital approaches to contact tracing that used location histories but many threatened individual rights and privacy [129]. The approaches we saw implemented

[1] I led this work and coauthored a paper that was published as a preprint with colleagues Michiel Bakker, Praneeth Vepakomma, Kent Larson, and Alex 'Sandy' Pentland. `https://arxiv.org/abs/2003.14412`.

were centralized, controlled by government authorities, or otherwise operated on a skewed trade-off between privacy and effectiveness. The goal of this work was to break past the false dichotomy of effective versus privacy-preserving contact tracing. We offered an alternative approach to assess and communicate users' exposure risk while preserving individual privacy. Our proposal uses recent GPS location histories, which are transformed and encrypted, and a private set intersection protocol to interface with a semi-trusted authority.

There were some other proposals for privacy-preserving contact tracing, based on Bluetooth and decentralization, that could further eliminate the need for trust in authority [112, 114, 113, 130, 131]. However, at the time the solutions with Bluetooth were technically limited[2] - Apple and Google had not yet announced any plans for their unified contact tracing framework. There was also the issue that Bluetooth-based systems required mass adoption before they could be effective, while location data was already being collected and used. Furthermore, these systems were not perfectly privacy-preserving either[3] and the solutions with additional measures for decentralization added additional complexity that hindered their viability.

The goal of this work was two-fold: To propose a location-based system that was more privacy-preserving than what was being adopted by governments around the world, and that was also practical to implement with the immediacy needed to stem a viral outbreak.

5.1.1 Trust & privacy principles

The privacy and trust principles central to the design of this work are summarized below:

- **Keep location data private.** Locations visited are kept private for all users including those who are diagnosed disease carriers.

- **Avoid surveillance.** The system can detect points of contact between users without precise location histories being exposed.

[2]Bluetooth-based systems suffered from interoperability issues across devices as well as iOS specific limitations on using Bluetooth in background processes.

[3]See a description for how these systems were susceptible to privacy attacks, namely by other users, in work by Cho et al. [126].

- **Only allow one-way private data publication.** Only diagnosed carriers ever publish data, but this data remains encrypted and private, and their identities remain protected. Other users can check if they came in contact with carriers without sharing their location histories.

Any privacy-preserving contact tracing framework should be considered a "best effort" and avoid promising to be perfectly private. Our primary contribution to the space of existing frameworks and digital tools was the degree to which our cryptographic approach could preserve user privacy while providing highly useful and accurate information through individuals' location data.

The following sections describe our proposed system design. The objective of this work was not to implement such a system. It was instead to show that there were effective and more privacy-preserving alternatives to the systems we saw governments adopting.

5.2 A GPS-based privacy-preserving scheme

5.2.1 A useful first step

A simple first version of a system that provides exposure risk information is one that collects, anonymizes, and aggregates the recent GPS location histories of diagnosed carriers. This information allows the creation of a spatiotemporal *heatmap* representing large geographic regions where diagnosed carriers spent time and when.

Individuals' data and areas visited must be aggregated and obfuscated in a way that minimizes what can be learned about individual people or places visited in the dataset. This is done in order to protect people and businesses from potential stigmatization or any other threat. This aggregated view can provide helpful information about infection risk across different areas, types of places, and time periods for both health authorities and the general public. This aggregated data can be further analyzed to better understand the flow and trends of disease transmission.

5.2.2 Contribution

This work builds upon this first step with a *private set intersection* protocol to provide more precise risk assessments to individual users based on points of contact with individuals who were later diagnosed as disease carriers. Our approach partitions the space of fine-grained GPS location and time data into discrete spatiotemporal points that represent location histories. This combination of a partition scheme and private set intersection protocol allows the system to detect when a user came in contact (e.g. was in close proximity) with diagnosed carriers to assess and inform them of their risk, while preserving the privacy of individuals.

The following sections describe what type of information our proposed system provides before showing a high-level system overview. I then explain the trust and privacy model it is designed for, and finally provide a more technical description with details on how the system could be built in practice.

5.2.3 Probabilistic risk assessment

The proposed system provides a probabilistic measure of disease exposure risk for a user, based on the time they have spent in spaces shared with users who were later diagnosed as disease carriers. More time spent in such shared spaces indicates higher levels of risk, but this risk is also dependent on where those spaces are.

Any technological system should be wary of claiming to precisely determine exposure risk, due to the limitations of the technologies used for detection, and the ambiguity over what types of interactions between people, shared spaces, or common surfaces, most elevate risk. Our proposed system uses GPS points collected from users' devices and can be extended to use Bluetooth to indicate locations visited as well. It is worth noting that GPS has limited accuracy, especially in dense urban environments or multi-story buildings. But even detecting whether a user spent sustained time in a crowd or multi-story building with a diagnosed carrier may be useful, due to the heightened likelihood of sharing not only space but surfaces such as door handles or elevator buttons, which help a virus spread [132]. For this reason, we proposed a probabilistic risk assessment based on the amount of time and

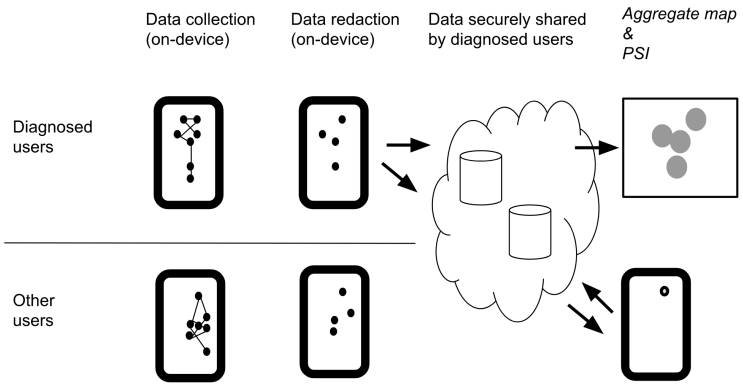


Figure 5-1: High-level schematic showing the major steps in the system's process for privacy-preserving contact tracing: (1) data collection, (2) redaction and transformation of data, (3) secure data exchange, and (4) individualized risk assessment and notification, as well as the distribution of aggregated data to create a 'heatmap' to inform the public of more general risk. A more detailed description of the process includes usage of discrete spatiotemporal 'point intervals' and a private set intersection protocol and is described later on in this chapter.

number of places that a user shared with a disease carrier, which we call "points of contact".

5.2.4 System overview

This section provides a high-level overview of our system and walks through an illustrated example.

Our proposed system follows four steps: (1) data collection, (2) redaction and transformation of data, (3) secure data exchange, and (4) risk assessment and notification.

Step 1: Data collection. An application (app), installed on the mobile device of the user, collects timestamped GPS points throughout the user's day, every t minutes. The sequence of points represents their location history.

Step 2: Redaction and transformation of data. All collected location histories are redacted, transformed, and encrypted before leaving the device, to protect user privacy. In the case that a user is diagnosed, their points are transformed and shared with the server in a way that maintains their privacy. Collected points are deleted from both devices and servers d days after they were collected, where d is the period of possible disease transmission informed by medical experts.

Step 3: Secure data exchange. In this phase, using the agreed secure data exchange mechanism, the mobile app (acting as the client) establishes a secure channel with the designated server. Within this secure channel, the mobile application requests the server for the 'point interval' data of known infected carriers for a chosen duration of time (e.g. the last 2 weeks) for a given region (e.g. Boston).

The infected carriers' data in the server's possession is anonymized and has already undergone redaction and transformation to remove sensitive information to limit the risk for re-identification, and contains no personally-identifying information (PII).

Users' apps can check whether they came in contact with carrier users, and how often, while preserving privacy. This is done with a cryptographically secure "private set intersection" (PSI) protocol to find matches between encrypted 'point intervals' for carriers and other users.

Step 4: Risk assessment and notification. The app assesses risk for its user based on the points of contact it found via (3) and can notify users of risk. Users whose apps find them at risk due to contact with carriers can then be encouraged to get tested or self-quarantine. The app can optionally show the user where and when the points of contact occurred.

Steps 2 and 3 are expanded upon in the following section.

Storing and sharing GPS histories

As GPS points are collected by a user's device, sensitive areas are removed through either automatic redaction or manually by the user. Redaction is an important privacy step, as knowledge about where someone was when, or where they commonly spend time, such as their home area, can be used to re-identify pseudo-anonymized users [104, 35].

The system provides two methods of redaction: automatic and manual. Home areas can be easily inferred by the app based on where users spend time in the nighttime, and these can be automatically redacted. In addition, the app can provide the user with an interface to mark additional sensitive areas for redaction. Any GPS point collected within an area marked for redaction is deleted and not shared. To further protect the user's privacy, this

redaction happens on the user's device and the remaining points are then transformed or obfuscated before they are stored. Redacting and modifying GPS points on the device, rather than after points are shared, is an important privacy measure to prevent users from being coerced into providing information on where they have been.

If a mobile app user is diagnosed as a carrier (e.g. by professional medical personnel), the proposed system provides two different ways for that user to anonymously share their GPS points to provide important information to healthcare professionals, other system users, and the general public.

The two GPS transformations that support these different use cases are:

(a) GPS points are replaced by larger geographic areas that contain them to represent the areas where carriers spent time and when, without representing precise locations.

(b) Precise timestamped GPS points are transformed into "point intervals" and obfuscated using a one-way hash function (e.g. NIST standard SHA256 hash algorithm).

The first way (a) is used for the aggregated view of data that was previously described as a motivating first step. The granularity level can be dialed-up or dialed-down depending on the circumstances. The more fine-grain granularity means an increase in likelihood that the user can be re-identified.

The second way (b) is used for contact tracing. This use case is where we made a new contribution with our approach to finding points of contact while preserving privacy.

The two different use cases that (a) and (b) serve are both central to this work. However the remainder of this chapter focuses on (b), as it is our main contribution and requires explanation.

5.2.5 Trust and privacy model

The proposed system is designed around a model that assumes there is a semi-trusted authority maintaining the server with diagnosed carriers' redacted location histories. Such semi-trusted authorities could be local hospitals or local government agencies chartered and

regulated to hold citizen data and maintain data privacy. We believe that a common goal – one that would make the proposed system usable while preserving individual privacy – is to minimize the amount of information from a diagnosed carrier that is exposed to other users and to the semi-trusted authority.

The proposed system is designed to minimize the amount of diagnosed carriers' information that is exposed, and to maximize the privacy for all other users of the system. These other users need not share any of their location data in order to find points of contact with diagnosed users. However, diagnosed users do risk forfeiting some privacy when they share their location histories with the authority managing the server [4]. Even though they only share their redacted and encrypted location data, given enough computational resources and malicious intent, the managing authority can attempt to circumvent these measures and reconstruct location history data from the encrypted data that was shared by the infected users[5].

There is a clear need for a *governance model* regarding this data collection and use. For example, data should be deleted d days after it was shared, where d is informed by medical experts. There should also be a legal framework in place to end the practice of collecting data in this way once the health crisis is under control[6].

That said, we also acknowledge that governments already have potential access to the massive amounts of location data that our proposed system would collect. Location histories are already collected by apps on users' smartphones, and the cellular towers they connect to, and through credit card purchases.

[4]Even systems designed to be opt-in on the part of users sharing data can be abused and made compulsory by authorities once they are built. In Singapore, people contacted by health authorities are required by law to assist in the activity mapping of their movements and interactions [112, 133].

[5]The authority managing the server with redacted and encrypted user data can attempt a grid search (brute force) attack over all possible encrypted point intervals in order to find hash collisions and reverse the one-way hash function that encrypted the point intervals that users shared, and thereby expose the underlying data. They can then attempt to reconstruct location histories based on the spatiotemporal correlation between data points or re-identify users due to the unique nature of location histories [134].

[6]We have examples to be wary of regarding measures taken in times of crisis that extend indefinitely. Israel declared a state of emergency during its 1948 War of Independence, justifying a range of "temporary" measures that removed individual freedoms. They won the War of Independence but never declared their state of emergency over, and many of the "temporary" measures are ongoing [135]. Israel also approved cellphone tracking for its COVID-19 patients [136]. Similarly COVID-19 surveillance innovations in China are likely to be used by China's counterterrorism forces beyond the pandemic to further monitor and regulate the movement of its people. Consider the Uighur people. The Chinese government has categorized this ethnic group as terrorists and has subjected many of them to forced labor [137].

5.3 Technical description

Our proposed system broadly involves data collection followed by a method to deterministically construct hashed spatiotemporal intervals that discrete points in users' location histories are mapped to. These intervals are then used with a private set intersection protocol to inform users when points in their location histories match the points in the location histories of diagnosed carriers. These steps are described in the following sections.

1. Collecting and Representing GPS points
2. Detecting Points of Contact Using Private Set Intersection
3. Assessing Risk and Notifying Users

5.3.1 Collecting and representing GPS points

Timestamped GPS points are collected within user devices as they move throughout their day. These points are collected as tuples of latitude, longitude, and time: (latitude, longitude, time). A user's app checks for matches between their collected points and the points shared by users who were diagnosed as carriers in order to identify points of contact.

Partition space and time into intervals

For privacy purposes, GPS points are never directly compared in order to find these matches. Timestamped GPS points are instead first mapped to a 3-dimensional grid, where two dimensions are for geographic space (latitude and longitude), and the third dimension is time. We call these 3-dimensional grid cells 'point intervals'. The point intervals are then obfuscated with a deterministic one-way hash function. Identifying points of contact becomes a matter of matching hashed point intervals. Transforming GPS histories in this way to map (latitude, longitude, time) points that occur within a continuous spatiotemporal space into discrete 'point intervals' makes comparing hashed GPS histories possible. It also makes sense for our use case of finding time intervals where users occupied the same spatial area[7].

[7]Phones collect GPS data with limited accuracy and therefore trying to match users across spatial areas with radii too small will miss points of contact.

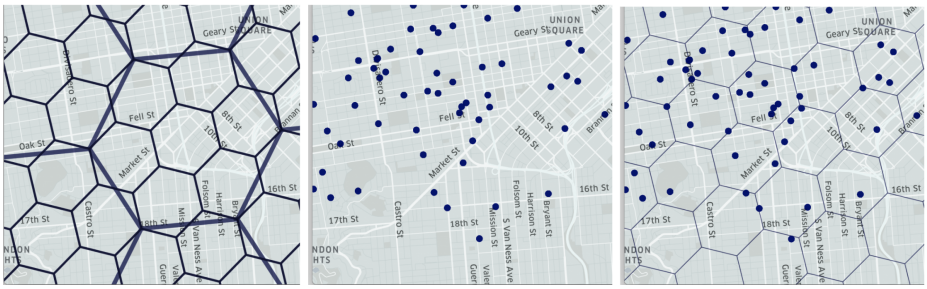


Figure 5-2: A geographic area partitioned by a hexagonal H3 grid. Points are mapped to an index corresponding to their containing grid cell.

There are established ways to partition a geographic space, such as with geohashes[8] for square grid cells or with the hexagonal global geospatial indexing system of H3 grid[9].

Similarly, time can be partitioned into intervals. For example, if an interval size is 2 minutes, then an interval boundary can always fall on the hour, and on the 2nd minute of the hour, and so on. The 3-dimensional grid of point intervals is an underlying system parameter (or logical "map") that is agreed-upon and shared across all user devices in the system. The specific partition scheme and interval sizes used are implementation details. What matters more is that the chosen partition scheme and the geographic and temporal resolution used are consistent across devices.

We note that when collecting GPS points there is a trade-off between accuracy in detecting points of contact and the amount of data that must be then stored and processed. For example, if data is collected more frequently, the system is more likely to detect when users spend little time near each other, such as sharing a bus ride. However, this requires collecting and storing more data, and hence more compute resources. We also note that the geographic partition of space can be expanded to include specific locations. For example, a bus line might install Bluetooth beacons on its buses with unique identifiers to serve as the geographic portion of point intervals, allowing users with an app that supports this functionality to later detect if they shared a bus ride with a diagnosed carrier.

[8] `https://en.wikipedia.org/wiki/Geohash`
[9] `https://h3geo.org/`

Checking for matches across obfuscated GPS histories

Since the point intervals are transformed with a one-way hash function, they cannot be easily reversed to expose underlying location histories of users.

Yet, since this transformation is deterministic, users' apps can still check for points of contact with diagnosed users by checking for matches between their transformed point intervals and those returned by a server. Verifying whether two individuals came into contact becomes a matter of comparing whether the hashed point intervals that were constructed from timestamped GPS points (latitude, longitude, time) collected by their devices coincide with any of the hashed point intervals provided by the server (as well as adjacent point intervals)[10].

5.3.2 Detecting points of contact using private set intersection

We now describe the protocol that facilitates the private detection of matches across users' GPS histories. We assume that there is a semi-trusted authority (e.g. a local health agency) operating a server. When a patient is diagnosed as a disease carrier, they share their redacted, anonymized, hashed point intervals to the central server. Other users' apps periodically exchange information about their own hashed point intervals with the server to detect if their hashed point intervals match against those shared by diagnosed carriers. They do so with a private set intersection protocol.

Private set intersection (PSI) enables two parties to compute the intersection of their data in a privacy-preserving way, such that only the common data values are revealed. It has applications in a variety of privacy sensitive settings, from measuring conversion rates for online advertising [138] to securely testing sequenced human genomes [139].

In our case, the two parties involved are the server storing the hashed point intervals shared by diagnosed carriers, and another user's device - the client. Their data are their respective

[10]Technical note about matching against adjacent intervals: Since the partition of geographic space into intervals was predetermined, two points may be close together in space but fall into different intervals. For this reason a user's app checks each of their collected point intervals, as well as the spatially adjacent intervals against the diagnosed carriers' shared point intervals. That is, if we use a spatial grid of hexagons (like H3), a user's collected point falls within a grid cell and that grid cell is used as an interval. We must check that interval as well as the surrounding 6 hexagons in the grid against each data point shared by diagnosed carriers. This adds some complexity in terms of processing power to make the comparisons $7N_u \times N_I$ rather than just $N_u \times N_I$.

hashed point intervals. We can leverage PSI in a way so that only the user learns about the intersection of their data - the server does not learn whether it shares any point intervals in common with the user, while the user's client app does learn this. Therefore, our use of PSI is designed to maximize the privacy for users who may be wary of surveillance or who do not fully trust the entities that maintain the server.

There are many PSI schemes that would fit our needs. These different schemes vary in their computational complexity, speed, and accuracy. Researchers have developed fast PSI protocols optimized for a client-server model, including those where the client is a smartphone app [140] and where the server's dataset is significantly larger than the client's [141], which is the case for our system[11]. A good overview and comparison of PSI protocols can be found in [138].

Example PSI scheme with Diffie-Hellman

To aid the reader in understanding how PSI supports our privacy goals, I provide a simple scheme using the Diffie-Hellman protocol [142, 143] in the appendix section B.2.

To briefly summarize the example: Infected users' point intervals are already obfuscated with a hash function and shared with the server. This set of hashed points intervals from infected users is represented as $H(P_I)$. Any other user's point intervals, P_U, are obfuscated with the same hash function to locally store $H(P_U)$. The goal with PSI is for a user's app to detect the intersection of the hashed point intervals, $H(P_I) \cap H(P_U)$, as this also then reflects the intersection of unhashed point intervals $P_I \cap P_U$. This detection is done locally within the app, without the app directly sharing the user's $H(P_U)$ with the server. In the example provided with a Diffie-Hellman approach, the user's app encrypts their hashed point intervals with a key, a, to send $H(P_U)^a$ to the server. The server encrypts $H(P_I)$ with its own key, b, to return the encrypted set $H(P_I)^b$. It also encrypts $H(P_U)^a$ with b to return $H(P_U)^{ab}$ as well. The client can then further encrypt $H(P_I)^b$ with a to result in the set $H(P_I)^{ba}$. Due to the multiplicative properties of the group under which the values were

[11] However many of these PSI protocols achieve their improved efficiency at the cost of accuracy, allowing a small number of false positives, such as by employing bloom filters. This may not be an acceptable trade-off for a disease contact tracing system where false positives can lead to panic or the wrong people seeking scarce medical resources.

encrypted, values in the intersection $H(P_I) \cap H(P_U)$ are also represented in the intersection $H(P_I)^{ba} \cap H(P_U)^{ab}$. The user's client app learns of this intersection while the server does not. The appendix section B.2 provides a more fully explained example which is also illustrated in figure B.2.1.

PSI benefits and implementation notes

The use of a PSI protocol adds an extra layer of privacy protection for both the diagnosed carriers and the other users, beyond just the redaction and obfuscation of data[12].

Other users only ever learn points shared by diagnosed carriers (P_I) that their own points (P_U) matched with (the intersection $P_I \cap P_U$). This further protects the privacy of diagnosed carriers. Moreover, the server need not learn whether any points match; only the other user learns the intersection of points. This further protects the privacy of undiagnosed users who may be wary of their location histories leaving their device, or being shared with an authority.

There are implementation details that can further enhance privacy and efficiency. For example, servers can hold data specific to local geographic regions, so that users with location histories specific to an area (e.g. the Boston area) need not interact with servers holding data specific to a far away region (e.g. the Bay Area in California). This helps subset the data so as to run PSI on a much smaller dataset, thereby helping computational efficiency. In addition, data shared by diagnosed carriers to servers should be deleted after d days, as both a privacy and efficiency measure.

The server should also limit the amount of data that a user's client device can exchange with it. It is only relevant to compare recent location histories (i.e. from the past d days). Since points are partitioned into consistent time intervals, there is therefore an upper bound on the number of points, N, that any app needs to check against the server's set of points, P_I. The server can limit the exchanges with any client to N points per exchange, and limit the number of queries per day. This limitation is important for preventing privacy attacks where

[12]When other users send their point intervals to the server, their point intervals are effectively encrypted twice. First with the deterministic hash function (H) that encrypts all point intervals in the same way. Second by the PSI protocol.

an adversary might attempt querying over the entire spatiotemporal grid to reconstruct the location histories of diagnosed carriers. It also reduces the computational burden for servers.

5.3.3 Assessing risk and notifying users

A user's app can assess their risk of infection based on the comparison (performed on the user's device) between the point intervals on the user's device with those received from the server. Users whose apps find them at reasonable risk can then be encouraged to get tested or self quarantine.

The implementation of our system can differ to either allow the app to learn just the number of points of contact that occured, or where and when points of contact occurred. These different implementations have different implications for the privacy and utility that our system can offer to its users.

The number of detected points of contact is related to how likely a user was to have spent sustained time in spaces shared with diagnosed users, so the number of detected points of contact is commensurate with risk and can be used to provide a risk assessment. When the locations of points of contact are known, the risk assessment can leverage context about these locations, such as whether they are confined spaces versus multi-story office buildings versus outdoor parks. Future work could further incorporate intelligence into the risk assessment.

5.4 Intermediary implementation

Given the urgency of the COVID-19 pandemic, we also described how intermediary steps could be taken to implement the presented system: Even before a secure server is set up to perform the private set intersection (PSI) protocol, hashed point intervals for diagnosed carriers can be published to a flat data file for other users to download. While this would speed up implementation, it would also diminish privacy guarantees[13]. Finally, this inter-

[13] Attackers could attempt to create points representing every potential point interval to check for matches. Attackers could then attempt to reconstruct location histories of users diagnosed as carriers and possibly re-identify them from their shared anonymized data. While the redaction step would decrease the likelihood of an attacker's success, some privacy risk remains.

mediary implementation with a flat data file could then be subsequently transitioned to the more secure implementation using a PSI protocol.

5.5 Discussion

In this work we proposed a technical design to address the problem of assessing users' risk of disease exposure with location histories. Our proposal was in response to existing digital contact tracing technologies, with a more privacy-preserving approach.

We also noted that in contrast to these other technologies, it was important for any implemented system to be opt-in, and to clearly communicate to users how it collects, retains, and uses data. This was in order to provide users the opportunity to weigh the trade-off between their individual privacy risk posed by sharing information with the system and the ongoing risk posed by the pandemic.

As this work was developed, we were also encouraged by other privacy-sensitive proposals that emerged for contact tracing [144, 126, 113, 114]. Some of these even extended our notions of privacy by removing the need for trust in authorities who might abuse their access to diagnosed patients' encrypted data and violate their privacy. However, these systems were more complex and required more infrastructure and coordination, making them more difficult to implement. The goal of the work presented in this chapter was to propose a system that was more privacy-preserving than the contact tracing technologies that we saw governments around the world adopting, but that could also be practically implemented with the immediacy needed to both stem the spread of disease and stem the adoption of *privacy-violating* technologies.

What was clear throughout this work was that contact tracing could be a highly effective strategy to mitigate the global health crisis, and location data collected from personal devices could serve as a powerful tool in aiding this effort. In the face of the COVID-19 pandemic, it seemed time for the ubiquitous collection of location data to serve as a tool for public health.

Chapter 6

Using location data to understand social distancing behavior: A New York case study

6.1 Introduction

The previous chapters discussed how location data *could be used to mitigate* the COVID-19 health crisis, namely to scale contact tracing efforts. This chapter and the following are about how location data *was used to better understand* the health crisis and its impacts.

Awareness that COVID-19 could wreak havoc on the United States medical systems grew in March of 2020. The United States declared a national emergency on March 13th. In response to the oncoming crisis, many local governments across the United States issued "stay-at-home" orders or shut down non-essential parts of their economies. These and other "social distancing" policies were implemented as strategies to reduce interpersonal contact and thereby limit disease transmission [145, 146, 147, 148, 149]. Even without these policies in place or enforced, many people independently practiced social distancing as a responsible behavior to "flatten the curve", or reduced their own activities out of personal health concerns.

During the same period, myself and a group of colleagues led by Esteban Moro gained access to an up-to-date, and highly granular location based services (LBS) dataset. The data set contained anonymized location history data collected from millions of mobile devices used in metropolitan areas across the United States.

The research opportunity and potential importance of working with such a dataset was clear. Location based services data provided unique opportunities to understand how people were practicing social distancing behaviors in the United States, the extent to which they were doing so, and how this differed across communities and demographics. This data would also be used to study the impact of those behaviors on disease transmission rates.

This chapter describes work done in collaboration with Esteban Moro, Michiel Bakker, and Matt Groh. In this work we used the LBS data to measure social distancing behaviors in the New York City metropolitan area. At the time, New York City was considered the epicenter of the pandemic in the U.S.

6.2 Data

Location based services data. The LBS data were provided by Cuebiq, a location intelligence and measurement company. They supplied anonymized records of GPS locations from users who opted-in to share their data. The data were shared with us under a strict contract with Cuebiq through their Data for Good program where they provide access to de-identified and privacy-enhanced mobility data for academic research and humanitarian initiatives. In order to preserve privacy, location data in the inferred residential and work areas for users were aggregated to the Census Block Group level, thereby allowing for demographic analysis while obfuscating the true home locations of the users.

The data were provided in the form of time-stamped coordinates reported by users' devices. From these coordinates we computed "stays" which represent visits to locations[1]. They were detected as clusters of coordinates where a user spent at least 5 minutes. From these "stays" we computed more metrics described below.

[1]We used the infostop algorithm [150]

We used other available data sources in combination with the LBS data to add additional insights.

POIs data. We used "points of interests" data, referred to as POIs, from Foursquare. This secondary data source provided information about public places people visited, such as names and locations, as well as categories (e.g. "grocery store" or "arts and entertainment"). This allowed us to study changes in social distancing behaviors and understand those changes, by the type of place where they occurred.

Census data. We inferred the home census areas (Census Block Group) for the LBS data users we studied based on where they reported locations in the nighttime hours. The American Community Survey (ACS) [103] reports estimated demographics for people living in each of these census areas. We used this data to link demographic information to the aggregation of the users based on their home census areas. This allowed us to study changes in mobility and behaviors across demographic groups.

Data panel. From the LBS data we selected a panel of users who were sufficiently active during our period of study [2] and who were residents of the New York City metropolitan area. We restricted our analysis to those users. This panel represents a sample of a sample of 567,000 people, or about 3% of the population. When we compared the inferred home census tracts for this population to the ACS 2018 population estimates [103], we observed a Pearson correlation of 0.68.

6.3 Metrics

From the LBS and computed "stays" data, we were able to derive a variety of metrics. We explored these as ways to better empirically measure and understand how social distancing behaviors were taking place.

[2]The data panel was restricted to users from whom there was location data reporting that they stayed in their home Census Block Group for more than 10 days during the period of February 17 to March 9.

These metrics included those traditionally used by researchers who study human mobility: Daily trips (measured as the number of times users had stays data at distinct locations outside their home area), daily distance traveled (measured as the line distance), and daily radius of gyration (this represents the size of the area covered by users [50]).

We also explored new, less traditional metrics which were more relevant to the COVID-19 pandemic and social distancing behaviors. For example, the daily number of trips to POIs by POI category, and the duration of the trips there, could help us understand which behavioral changes were occurring. Before we decided to focus our initial analysis on the New York area, we were considering analyses across the United States. We then considered measuring trips by POI category as a means to quickly infer where and when local governments were enacting social distancing policies. For example, by detecting the municipalities with significant reduction in visits to schools or public offices, we might detect local policy changes made across the U.S. in an efficient way when otherwise that information was difficult to quickly attain.

We considered many other metrics, such as the density of users in public spaces, or the diversity of places visited by users on a daily basis. In each case, the LBS data source provided a unique opportunity to derive these metrics. The most important insight in the exploration of new metrics was that we could measure what the following work calls "contacts". The "contacts" metric counts instances when an individual user comes into proximity with another user. We were able to detect such instances with the LBS data and attribute them to precise locations and durations of occurrence, as well as count and distinguish the instances across users. Given that the primary goal of social distancing was to limit disease transmission through interpersonal contact, this seemed like the most relevant metric.

The following work focuses on these primary metrics and their relative changes:

Mobility

- Distance traveled.

- Radius of gyration.

- Number of people staying home.
- Number of stays in public places, which this chapter calls visits.

Contacts

This metric was estimated as instances where two people stayed within 25 meters distance from one another for at least 5 minutes.

6.4 NY Case study: Initial findings

We published our initial findings in a report at the end of March, which we updated mid April[3] [4] [151]. The report also contains more details about our methods used.

Here I provide a summary of some of our main findings along with figures illustrating them. The figures are timelines showing changes in behavior and the areas in gray are weekends.

In general the data showed dramatic changes in where people spent their time and with how many people they interacted following the declaration of the national emergency and implementation of social distancing policies. For example, when comparing weekends in February and late March, we found:

- Distance travelled dropped by 70% from a weekend average of 25 miles in February to 7 miles in March.
- The number of contacts in places decreased by 93% from 75 to 5.
- The number of people staying home the whole day increased from 20% to 60%.

The data also showed that in normal times, mobility and social contacts vary significantly by the demographic composition of a neighborhood. We studied this by attributing demographic information from the ACS by census tract of the residents in our data panel.

[3] Report "Effect of social distancing measures in the New York City metropolitan area": `http://curveflattening.media.mit.edu/socialdistance4_14.pdf`.

[4] The report reflects work led by my colleagues that I contributed to.

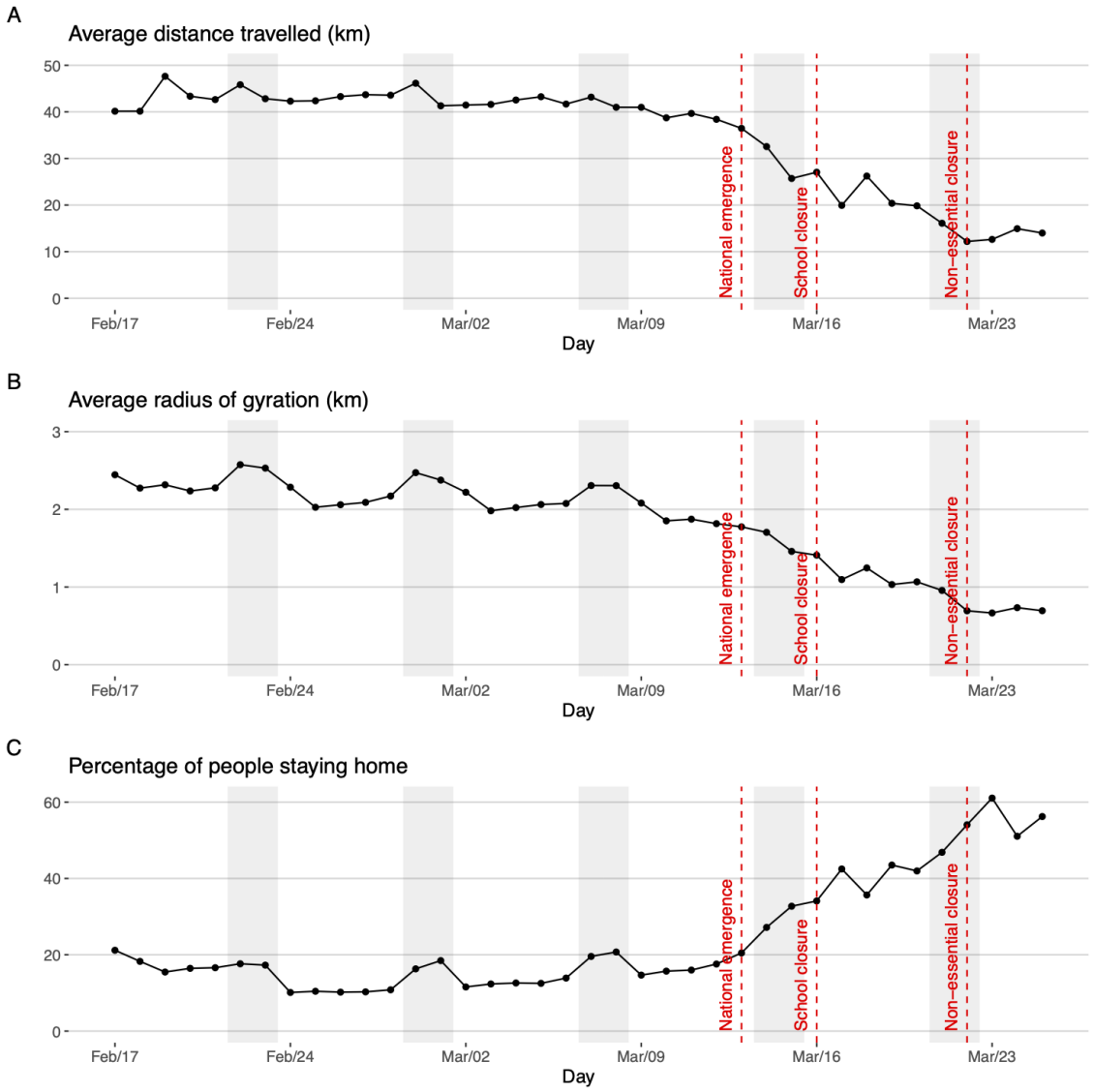


Figure 6-1: Changes in mobility metrics in the New York City metro area.

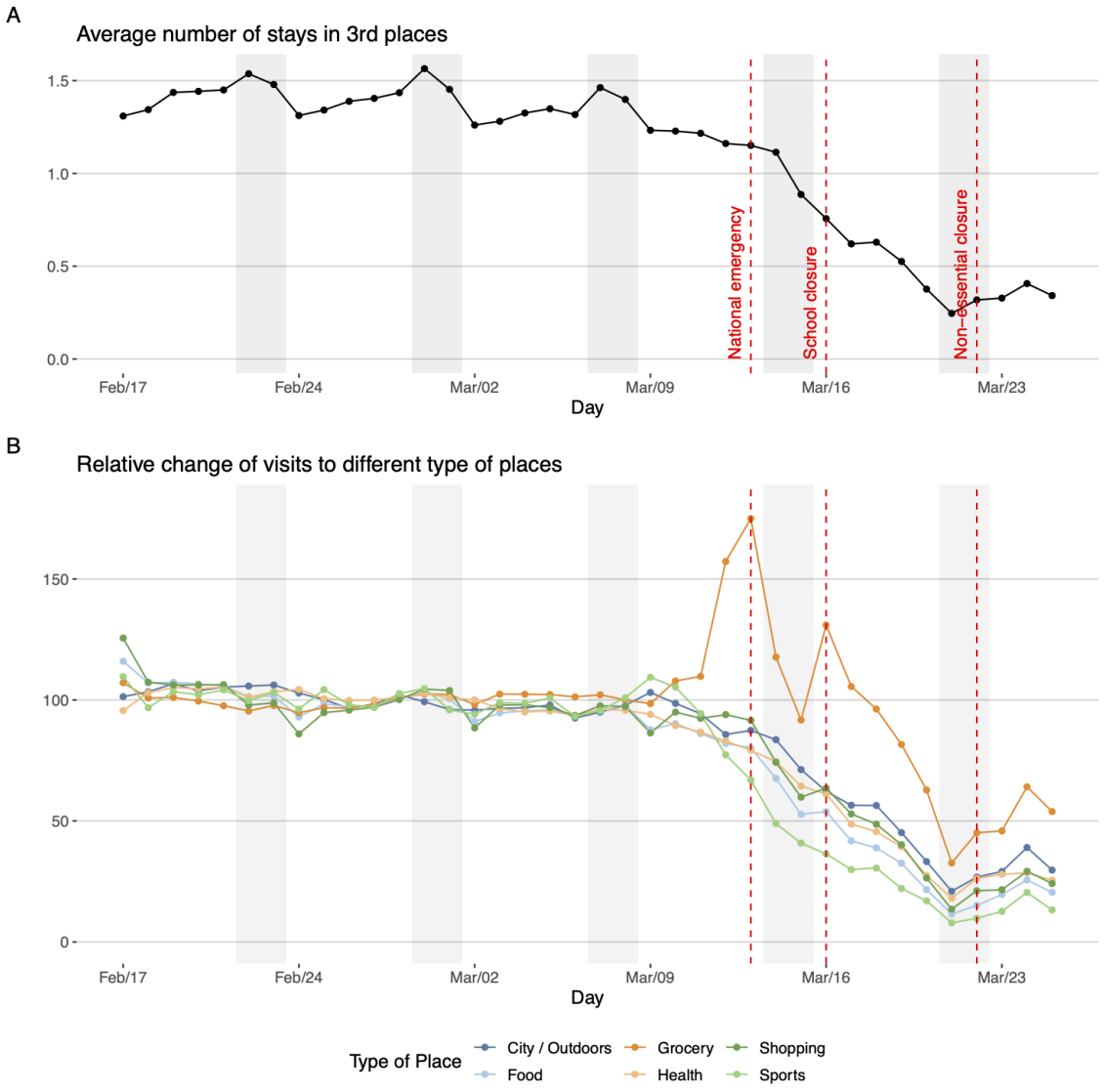


Figure 6-2: Change in visits to public places (i.e. stays in 3rd places).

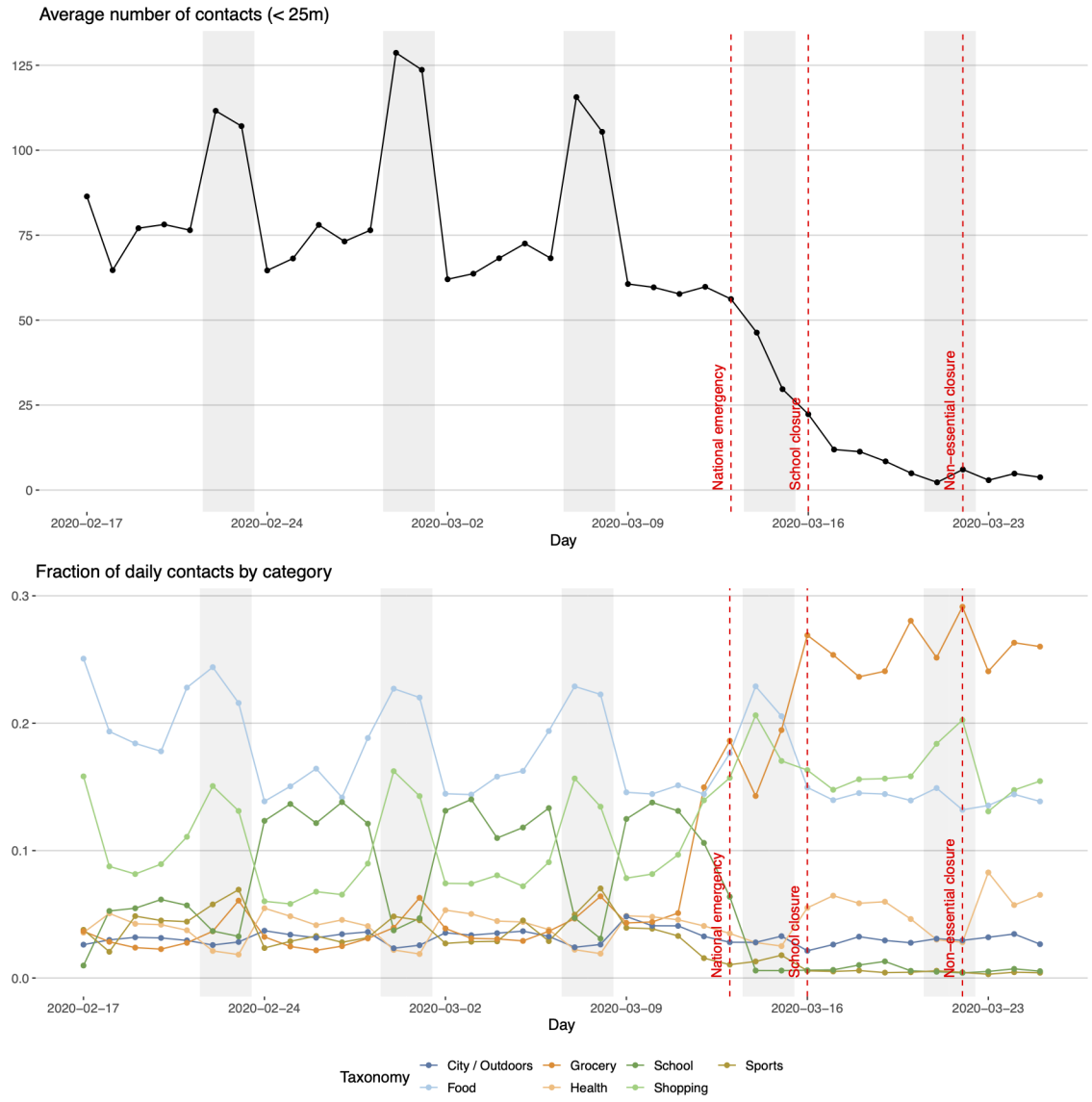


Figure 6-3: Change in the number of average daily contacts in the New York City area. The bottom figure shows the fraction of where these contacts occur by POI category.

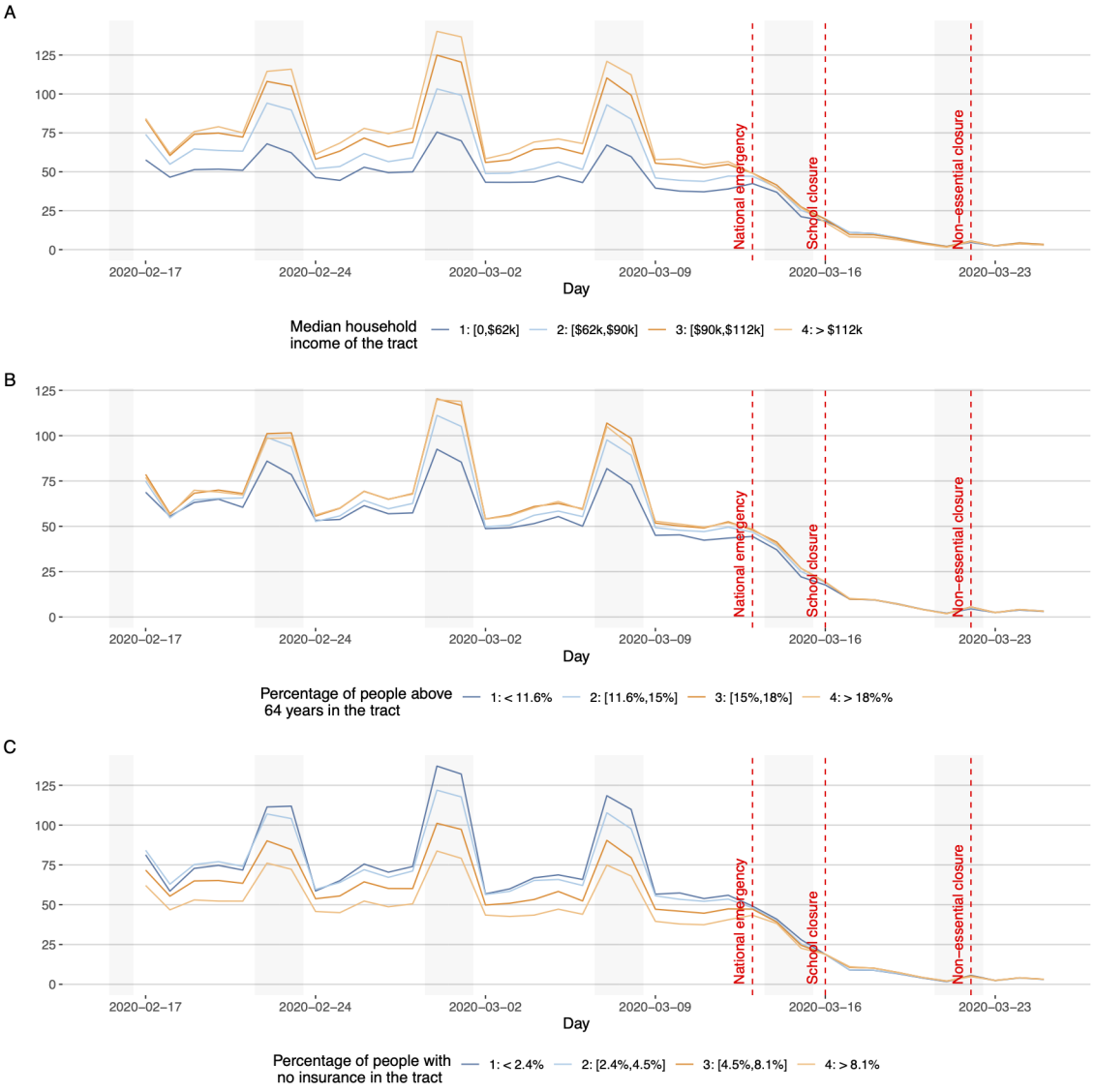


Figure 6-4: Average number of contacts by day by people in different groups of tracts. A) By median income, B) by percentage of population in the tract above 64 years old, and C) by percentage of people with no insurance.

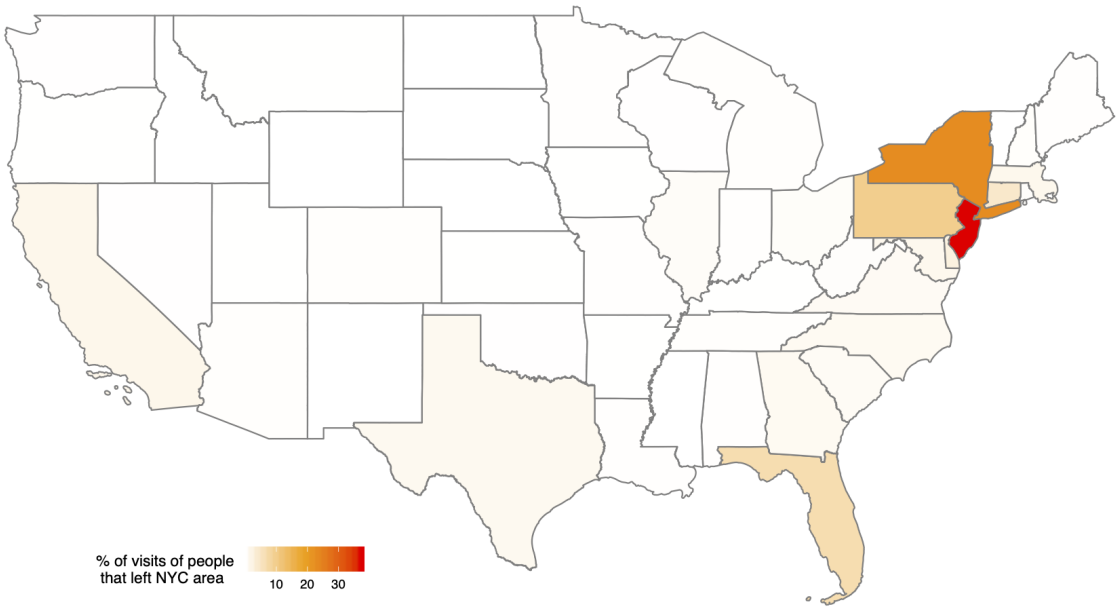


Figure 6-5: Percentage of visits to public places outside the New York City metro area during the weekend of February 21, from people in our data panel.

Following the social distancing policies, nearly everyone's mobility and social contacts were dramatically reduced to similar levels. This is shown in figure 6-4.

We also found that there was a spike in trips to grocery stores following the declaration of the national emergency (i.e. panic buying). Even after this time, grocery stores continued to be the primary places that New York residents came into contact with others (see figures 6-2 and 6-3).

In addition, we were able to measure the rate at which people (users in our data panel) were leaving the New York City area for other states (shown in figure 6-5). This was concerning from a public health perspective, as it was possible that these people were bringing coronavirus from the epicenter of the crisis to these other states.

6.5 Further work

This report, and the LBS data it used, provided real-time information about how social distancing was taking place in New York. Following the report, we continued using the data

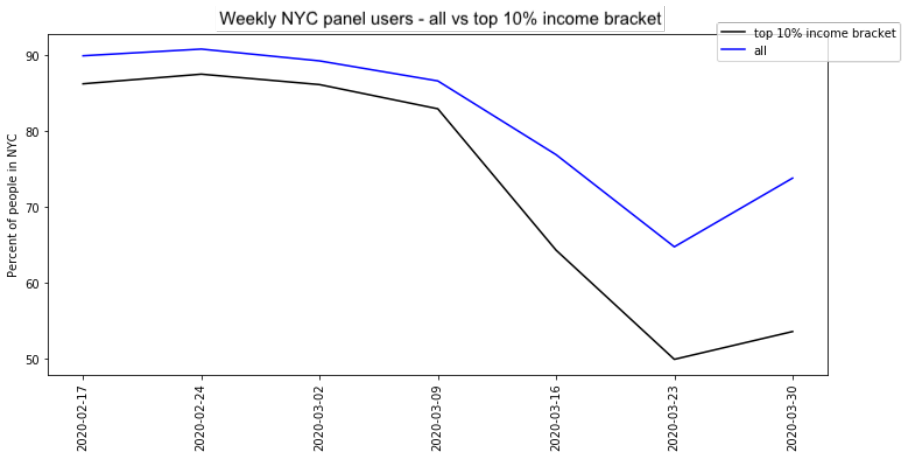


Figure 6-6: Percent of New York City residents in our data panel who left the city by week, compared by income bracket.

source to derive additional insights.

In particular, we were interested in how social distancing measures impacted different income groups. We studied the behaviors of panel users by the relative income brackets of their home census tract. For example, there was a spike in trips to grocery stores directly following the declaration of the national emergency on March 13. We found that while this change in behavior was significant for users in the bottom 10% income bracket, this was not at all the case for users in the top 10% income bracket.

We also looked at the disparate opportunities to escape the epicenter of the pandemic. We knew anecdotally that many people who could leave New York City, such as for parents' homes or second homes, did. This option was a privilege. Our data helped us quantify it: we measured how many users in each income bracket in our starting data panel continued to report data in the New York area each week. We found that those residing in the top 10% income bracket areas left New York in much greater rates than other New Yorkers (see figure 6-6). This kind of analysis was later independently done and reported on by the New York Times in their coverage of how the pandemic was impacting different socioeconomic groups[5].

There were many other questions about people's behavior in the time of the pandemic that we used our data to answer. For example, did special events, like Easter, lead to

[5] `https://www.nytimes.com/interactive/2020/05/15/upshot/who-left-new-york-coronavirus.html`

significant increases in mobility and interpersonal contacts, which could increase disease transmission? (For Easter the answer was no.) The LBS data source provided empirical answers to questions that may have otherwise relied on anecdotal information. The data also helped us fact-check claims made from other sources, such as by attempting to reproduce results from other researchers. Other researchers and news outlets had access to this data and similar datasets as well; overall it contributed important information to better understand the health crisis at a time of panic and uncertainty.

The nature of this data, due to how it is collected from personal devices, made it particularly informative. For example, its real-time nature delivered timely information. The fact that we could trace users across the dataset allowed us to infer home areas, infer demographic information, and glean additional insights. The high granularity of the data provided nuanced insights as well. It allowed us to study the types of places (POIs) where behavioral changes took place. More importantly, it allowed us to detect and measure interpersonal contacts.

However, while our analysis leveraged the highly precise nature of the data, it preserved the privacy of individuals. From the high-precision data we derived the mobility and contacts metrics, and then studied them in aggregate. For example, we were able to count the number of daily contacts at grocery stores by the census tracts for users who made the contacts, rather than for any individual users.

The use of aggregate contacts metrics provides a motivating example for how high-precision LBS data can be highly useful in privacy-preserving ways, particularly in service of a public health crisis.

Aggregate mobility metrics, such as the number of daily trips or distance traveled, by census area, are often used. We can imagine common use of an aggregated contacts metric as well. This metric could provide a better estimation for social distancing behavior than the more traditional mobility metrics that we saw other researchers use and news organizations report on during this time [152]. We also saw researchers use aggregate trips and distance traveled metrics as proxies for social contact in disease transmission models [153, 2, 154, 155]. Contacts metrics could better serve these models as well.

Future work could further leverage LBS data, or similar data sources, to derive contacts metrics in different ways. For example, we could measure the diversity of contacts, as the number of unique contacts a person has each day, or count contacts by people from different census tracts or communities. Or we might estimate the duration of contacts. Each of these differences may have different implications for measuring social distancing behaviors and potential disease transmission. And each of these differences can provide for different sources and uses of aggregate contacts metrics.

The potential utility of contacts metrics in addressing the health crisis motivates the work in the following chapter.

Chapter 7

A metric to better understand social distancing: Contacts

This chapter extends the work described in the previous chapter to use our "contacts" metric to address the monitoring and modeling of infectious disease transmission. This work was done in collaboration with Bernardo Garcia Bulle, Esteban Moro, and Michiel Bakker [1].

7.1 Introduction

In the beginning of 2019 the outbreak of COVID-19 spread rapidly across the United States as well as the rest of the world. In the absence of a vaccine and limited medical resources to meet the demands of the disease, many local governments throughout the U.S. issued stay-at-home orders or otherwise encouraged minimizing person-to-person contact, known as "social distancing".

Early studies indicated that social distancing interventions could be effective, such as the severe travel restrictions in China [156, 157] and stay at home orders issued both abroad [158] and in the U.S. [159]. Some of these studies measured the correlation between the time public interventions took effect and the change in growth of reported cases [160].

[1] I led the work in this chapter and produced its contents.

Many other studies used mobility metrics in order to monitor and model how social distancing measures impacted human mobility and disease transmission. These mobility metrics were either collected by surveys or computed from geolocation data collected from smartphones. These uses of geolocation data and mobility metrics are the focus of our work. For example Lai et al. used daily aggregate mobility metrics obtained from Baidu location based services to model the impact of travel restrictions in China with an SEIR framework [153]. Similarly in the U.S., Pei et al. used SEIR models to estimate the difference in likely COVID-19 deaths if U.S. lockdowns had occurred on earlier dates. Their models used census data reporting inter-county commuting flows from 2015 (ACS) in combination with data about visits to points of interest (POIs)[2] [2]. Many other studies in the U.S. used mobility metrics from geolocation data to quantify the extent to which social distancing occurred due to policy changes [161, 162] as well as measure and model the impact of social distancing on the growth of reported COVID-19 cases [154, 155].

A critical limitation of these works is their use of mobility metrics as *proxies* for the extent to which people were social distancing and limiting their *contacts* with others.

Contribution. We measure contacts by leveraging highly granular anonymized geolocation data to detect when people are co-located for an extended duration of time. This "contacts" metric intuitively serves as a more direct proxy for social distancing behavior and exposure risk, as limiting contact between individuals is the intention of social distancing measures. Our work uses location data collected from more than 1.8 million personal devices from 7 U.S. metropolitan areas.

We show the relationship between our contacts metric and mobility metrics used by other researchers, namely trips and distance traveled. We do so with analytical theory and empirical results.

This serves 2 main purposes.

First, we show how our daily "contacts" metric can be estimated by the more common

[2]Despite the limits of their data source, estimates from this model were featured in a New York Times article titled "Lockdown Delays Cost at Least 36,000 Lives, Data Show", published on May 20, 2020: `https://www.nytimes.com/2020/05/20/us/coronavirus-distancing-deaths.html`.

mobility metrics - either trips or distance traveled. This could improve the monitoring of local changes in social distancing and potential disease transmission.

Second, and more importantly, we show how the relationship between contacts and these other mobility metrics *changed* over the course of the health crisis. This is not surprising, as people changed their behaviors to adapt to social distancing. In particular, we find that increases in daily trips and distance traveled metrics do not correspond to increases in daily contacts as much as they previously did. This can be seen in figures 7-1 and 7-2. They show time series plots for the daily average metrics for each of the metropolitan areas in our dataset.

The change in the relationship between metrics that we identify signals weaknesses in relying on these other mobility metrics as proxies for social distancing behaviors. Our "contacts" metric might serve as a better proxy and it could be used as a daily aggregate metric in ways similar to how the trips and distance metrics currently do.

We also study the relationship between these mobility metrics and growth in reported cases for each of the 7 metropolitan areas in our study, localized to county. In particular, this work casts doubt on whether any of the metrics serve as sufficient proxies for disease transmission models in the U.S. the ways researchers intend for them to.

Outline. In the following sections we first show the relationship between the contacts metric and trips and distance traveled metrics with a theoretical framework (7.2), which we later validate in section 7.4. We describe our data sources and how we compute the metrics that are used in our analysis in section 7.3. We then show our analysis methods and results in section 7.4.

7.2 Relationship between contacts, trips, and distance metrics: Theoretical framework

Previous works have addressed the scaling relationship between population density and the rate of contacts which can lead to infectious disease [163, 164, 165]. This section builds

upon these works to show how the contacts metric can be estimated by the more commonly available mobility metrics of daily trips and distance traveled. We describe theory that we later validate with regression models in section 7.4.

We consider (and later compute) these metrics localized to geographic areas, namely U.S. counties.

- $contacts(t)$: The total number of times individuals from a given area come into contact with others, outside their own homes, on day t.
- $dist(t)$: The total distance traveled by individuals from a given area on day t.
- $trips(t)$: The total number of trips taken by individuals from a given area on day t.

We form the following relationships between these metrics:

$$contacts_i(t) \sim trips_i(t)^2 \tag{1}$$

$$contacts_i(t) \sim [populationDensity_i^{\alpha} \times dist_i(t)]^2 \tag{2}$$

For area i, where $populationDensity_i$ is the population density for area i, and where α is a positive value less than 1.

There is a third relationship between distance and trips (3).

$$dist_i(t) \sim \left[\frac{1}{populationDensity_i}\right]^{\alpha} \times trips_i(t) \tag{3}$$

Relationship (3) can be reformulated as (4)

$$trips_i(t) \sim populationDensity_i^{\alpha} \times dist_i(t) \tag{4}$$

and relationship (2) is a result of combining relationships (1) and (4).

In what follows we explain the theory for relationships (1) and (3), from which the other relationships follow.

7.2.1 Relationship between distance traveled and trips

We might intuitively assume that users in more urban areas travel shorter distances on average. For example, a city resident may not travel far to work or the grocery store, while a suburban resident outside a city does. Previous works studying the relationship between urbanity and average trip distance have shown this to be the case [166, 167, 168]. Population density is a commonly available metric. In the following theory and analysis we use population density as a proxy for an area's urbanity (which other works have done as well [167]).

Relationship (3) states our theory that the average distance traveled by users in an area depends on the average number of trips, weighted by the population density of that area,

$$dist_i(t) \sim \left[\frac{1}{populationDensity_i}\right]^{\alpha} \times trips_i(t) \tag{3}$$

where a positive α value means users in less dense areas travel further to reach their trip destinations.

In section 7.4 we use a log-linear regression model to validate this theory and identify the α values for the counties and metro areas in our dataset. We find consistent α values that are positive but less than 1, indicating that the marginal impact of population density on average distance traveled diminishes as population density increases.

7.2.2 Relationship between contacts and trips

The average number of contacts on day t for a given area is proportional to the square number of people taking trips on day t.

$$contacts_i(t) \sim trips_i(t)^2 \tag{1}$$

Contacts that increase exposure risk can occur when two users are at the same place at the same time. Ignoring contacts that occur within the home, users must make trips to these places where they come into contact.

Consider two average users, u_j and u_k at time t. We approximate the likelihood of them coming into contact based on the probability of them each making a trip versus staying at home, and whether their trips are to the same destination.

$$\begin{aligned} contact_{j,k}(t) = & \; 0 \times P(u_j \, stays \, home) \\ & + 0 \times P(u_k \, stays \, home) \\ & + a \times P(u_j \, makes \, trip \, at \, time \, t) \; \times P(u_k \, makes \, trip \, at \, time \, t) \end{aligned}$$

where a is some constant representing the average likelihood that users make a trip to the same place.

For any average user, u_j, we can then approximate their number of contacts, $contacts_j(t)$, by summing this approximation over all other users.

$$contact_j(t) \sim P(u_j \, makes \, trip \, at \, time \, t) \; \times \sum_k P(u_k \, makes \, trip \, at \, time \, t)$$

Since $trips(t)$ approximates $\sum P(u \, makes \, trip \, at \, time \, t)$ we can reformulate this as

$$contact_j(t) \sim P(u_j \, makes \, trip \, at \, time \, t) \times trips(t)$$

When we sum over all users u_j this becomes

$$\begin{aligned} contacts(t) & \sim \sum_j P(u_j \, makes \, trip \, at \, time \, t) \times trips(t) \\ & \sim trips(t) \times trips(t) \\ & \sim trips(t)^2 \end{aligned}$$

In section 7.4 we empirically validate this theory with log-linear regressions where the exponent, which our theory indicates as 2, is a free variable.

7.3 Data and computation of metrics

This work uses two primary data sources: geolocation data and daily reported cases. The following sections describe the data sources and the computation of the following metrics which we use in our analysis.

- $contacts_i(t)$: The number of contacts by residents of county i on day t.
- $trips_i(t)$: The total trips made by residents of county i on day t.
- $dist_i(t)$: The total distance traveled by residents of county i on day t.

7.3.1 Geolocation data

Data source and privacy

Contacts and mobility metrics are calculated by using data from Cuebiq, which is a location intelligence and measurement company. The data was provided as anonymized GPS locations from users who opted-in to share their data anonymously through a GDPR-compliant framework. Cuebiq provided the data through their Data for Good Program, where they provide access to de-identified and privacy-enhanced mobility data for academic research and humanitarian initiatives only. All researchers followed a strict contract obligating them to not share data further or to attempt to de-identify data.

Data panel

The analysis is limited to data from users whom we determined provided sufficient data to infer their area of residence. Specifically, it includes data from people for whom there is location data reporting that they stayed in their home Census Block Group more than 10

days during the period of February 17 to March 9. The final data panel includes 1,826,382 total users. See figure C.1.1 and table C.2 in the appendix C.1 for more information on panel users and data representativeness.

7.3.2 Reported cases data

Reported COVID-19 cases are provided at the county-level by the New York Times as daily cumulative diagnoses [169][3]. A limitation of this dataset is that it only includes reported diagnoses rather than true infection rates, and access to testing may vary across counties.

7.3.3 Geography and time frame

Our analysis uses data from counties surrounding 7 metropolitan areas: New York City (23 counties), Washington D.C. (24 counties), Dallas (13 counties), Boston (7 counties), Seattle (3 counties), Miami (3 counties) and LA (2 counties). The geolocation data used was collected between February 17 2020 and June 27 2020. This was the data available to us.

7.3.4 Computation of mobility and contacts metrics

The data are provided as time-stamped GPS coordinates sent as "pings" from user devices. The first part of computing the contacts and mobility metrics is using the data to compute likely "stay" points that best represent clustered locations reported by user devices. This computation is done using the Infostop algorithm [150]. For each of the visits to a location, a stay is the time a user arrived at a location, the duration of the stay, and the median latitude and longitude computed for that location. The minimum duration of a stay is 5 minutes and each stay consists of at least 2 pings in order to avoid creating stay points that represent when people walk or drive past a location.

We then compute the contacts metrics by detecting when two users have a stay within 25

[3]New York Times COVID-19 data are provided open source via Github: `https://github.com/nytimes/covid-19-data`.

meters distance from one another for at least 5 minutes[4].

The $contacts_i(t)$ metric used in our analyses counts the number of times a user with residence in county i comes into contact with another user. We only count contacts that occur outside users' home areas (census block). When both users are residents of county i, $contacts_i(t)$ is incremented for both of them.

The trips metric for each county, $trips_i(t)$, is the total number of stays users from county i have outside their home areas.

The distance metric is a proxy for the total distance people travel in a day. $dist_i(t)$ is measured as the total is the total line distance between consecutive stay points for users with residence in county i. Consecutive stay points within users' home areas do not add to this metric.

7.4 Analysis and results

In our results we focus on the change in the relationship between the trips and contacts metrics. However, we note that the trips metric is highly correlated to other mobility metrics such as distance traveled and the number of users staying at home. This is shown in the following sections.

Figures 7-1 and 7-2 show time series plots for average daily contacts and trips metrics for each metropolitan area in our dataset. These plots more simply illustrate the trends and changes that our analysis results describe numerically.

These average daily metrics are computed over all users in the panel. However there is variation in the number of daily trips and contacts for each user. The distributions of these per user metrics are shown in a series of plots in section C.2 in the appendix. In general the distributions are highly skewed towards a small number of trips and contacts per user and the distributions have long tails.

[4]Detecting contacts was done via an exhaustive search where each stay point was compared to each other stay point for other users in the same metropolitan area. This method is not computationally efficient.

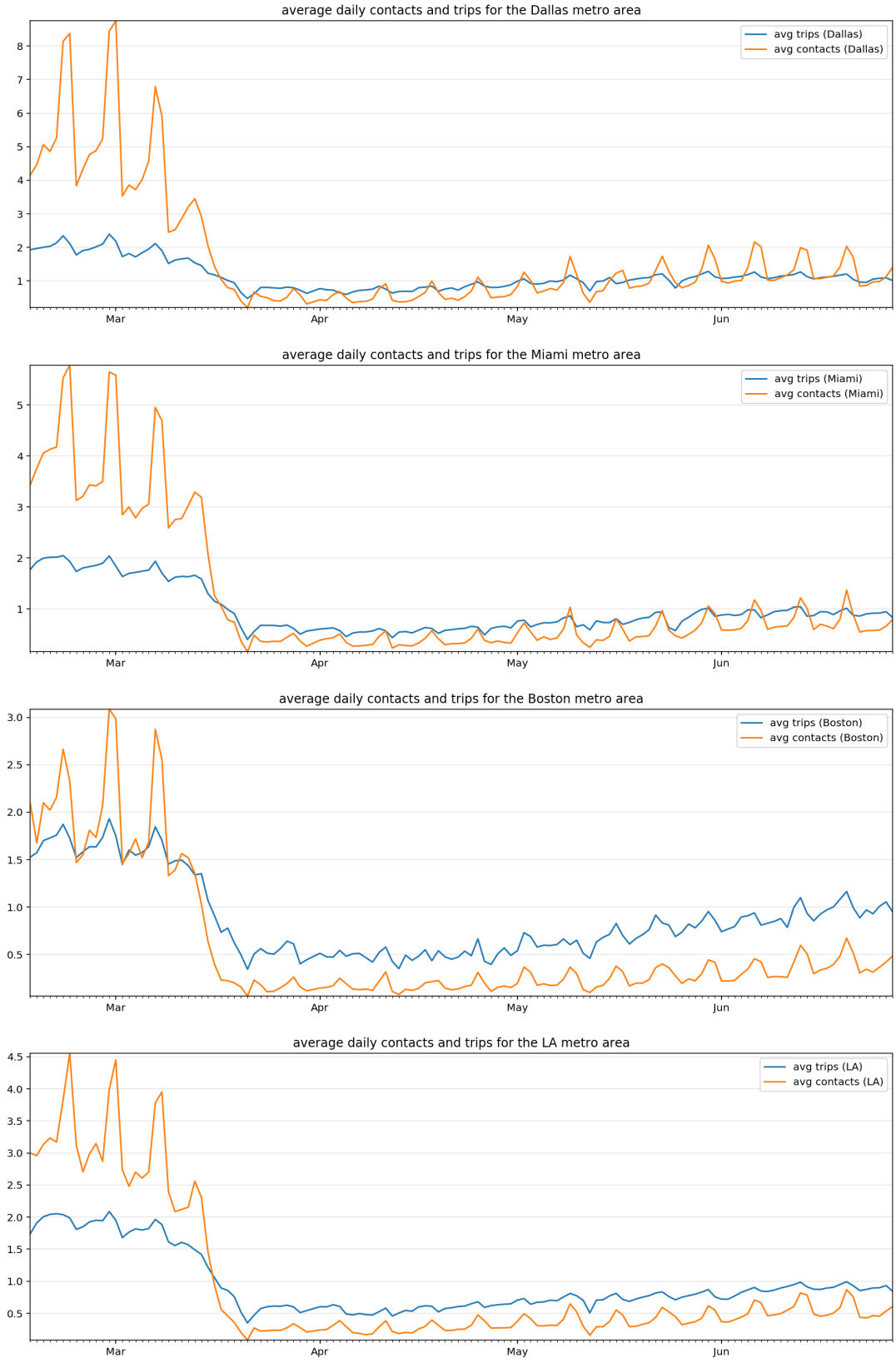


Figure 7-1: Time series plots for daily average trips and contacts.

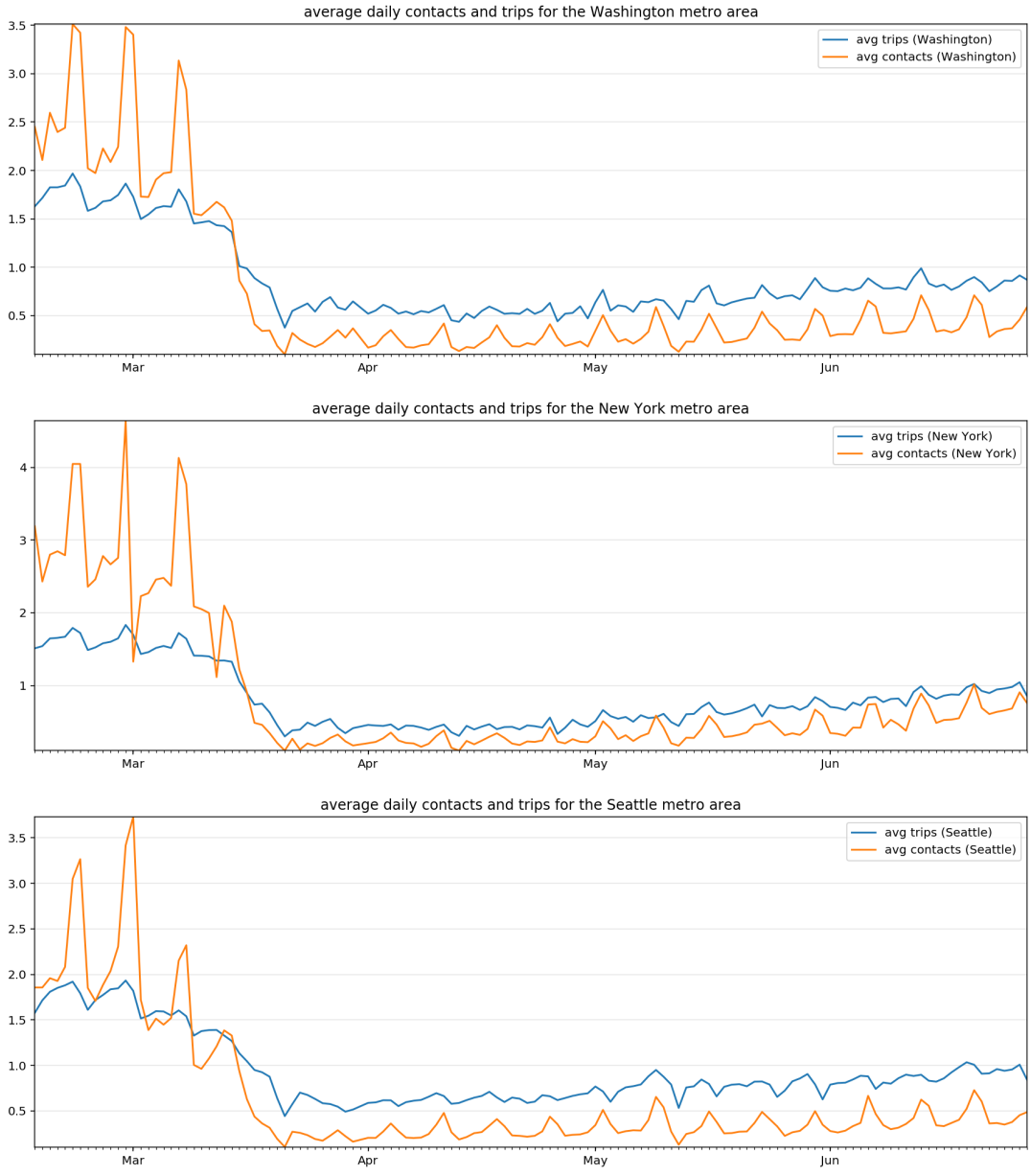


Figure 7-2: Continued: Time series plots for daily average trips and contacts.

7.4.1 Relationship between users staying at home and trips

Social distancing measures are often framed as "stay-at-home orders" intended to cause more people to stay at home. In addition to our main analysis and results, we estimate the relationship between the number of daily users staying at home and daily trips. The number of users who are *not* staying home scales linearly with the number of trips.

$$trips \sim [1 - (\text{portion of users staying home})]$$

See the appendix section C.2 for details about our model and the results in figure C.2.6.

7.4.2 Relationship between trips and distance traveled metrics

Our theory about the relationship between daily trips and distance traveled for a given area is described by the equivalent relationships (3) and (4).

$$dist_i(t) \sim \left[\frac{1}{populationDensity_i}\right]^{\alpha} \times trips_i(t) \tag{3}$$

$$trips_i(t) \sim populationDensity_i^{\alpha} \times dist_i(t) \tag{4}$$

We estimate α with a log-linear regression. The analysis is done for each metro area separately, using the daily trips and distance traveled metrics for each county, i, within the metro area. Population density for each county is calculated using census data from the ACS 2018 [103].

Our results are consistent across metro areas with a value for α between 0 and 1. See the appendix and table C.1 for details on the regression. The positive value for α is consistent with the hypothesis that lower population densities result in greater distances traveled for each trip. We interpret the value of less than 1 to mean that the marginal impact of population density on average distance traveled diminishes as population density increases.

7.4.3 Relationship between trips and contacts[5]

Figure 7-3 shows daily average trips versus daily average contacts for each county, colored by metro area, and where metrics are normalized by the county panel size.

We use linear regressions to empirically evaluate the theory that data should fit the model of relationship 1.

$$contacts_i(t) \sim trips_i(t)^2 \tag{1}$$

Specifically, variables a and b in the model

$$contacts_i(t) = a \times trips_i(t)^b$$

are estimated as free variables with the hypothesis that b is 2. This is done using the linear regression model $\log(contacts_i(t)) = \log(a) + b \times \log(trips_i(t))$.

The empirical results are generally consistent with our hypothesis. Using contacts and trips data computed at the metro level over the entire period of our data (February 17 to June 27), we estimate $b = 2.08$ (95% CI 2.03 to 2.13). See figure 7-4.

Our notable finding is that the relationship between contacts and trips changed with social distancing measures, especially following the U.S. declared a national emergency due to COVID-19 on March 13.

Figures 7-1 and 7-2 show this: there was a significant drop in each of the trips and contacts metrics directly following the emergency declaration.

More importantly - later there were increases in daily trips (and distance traveled) without commensurate increases in daily contacts. In other words, the relationship between the mobility and contacts metrics changed over time.

We numerically estimate the change by fixing $b = 2$ and estimating the change in a. The

[5] Code notebook that produced the plots and results in this section: `https://github.com/aberke/covid-19/blob/master/contacts_v_mobility.ipynb`. The notebook has additional county and time specific plots as well.

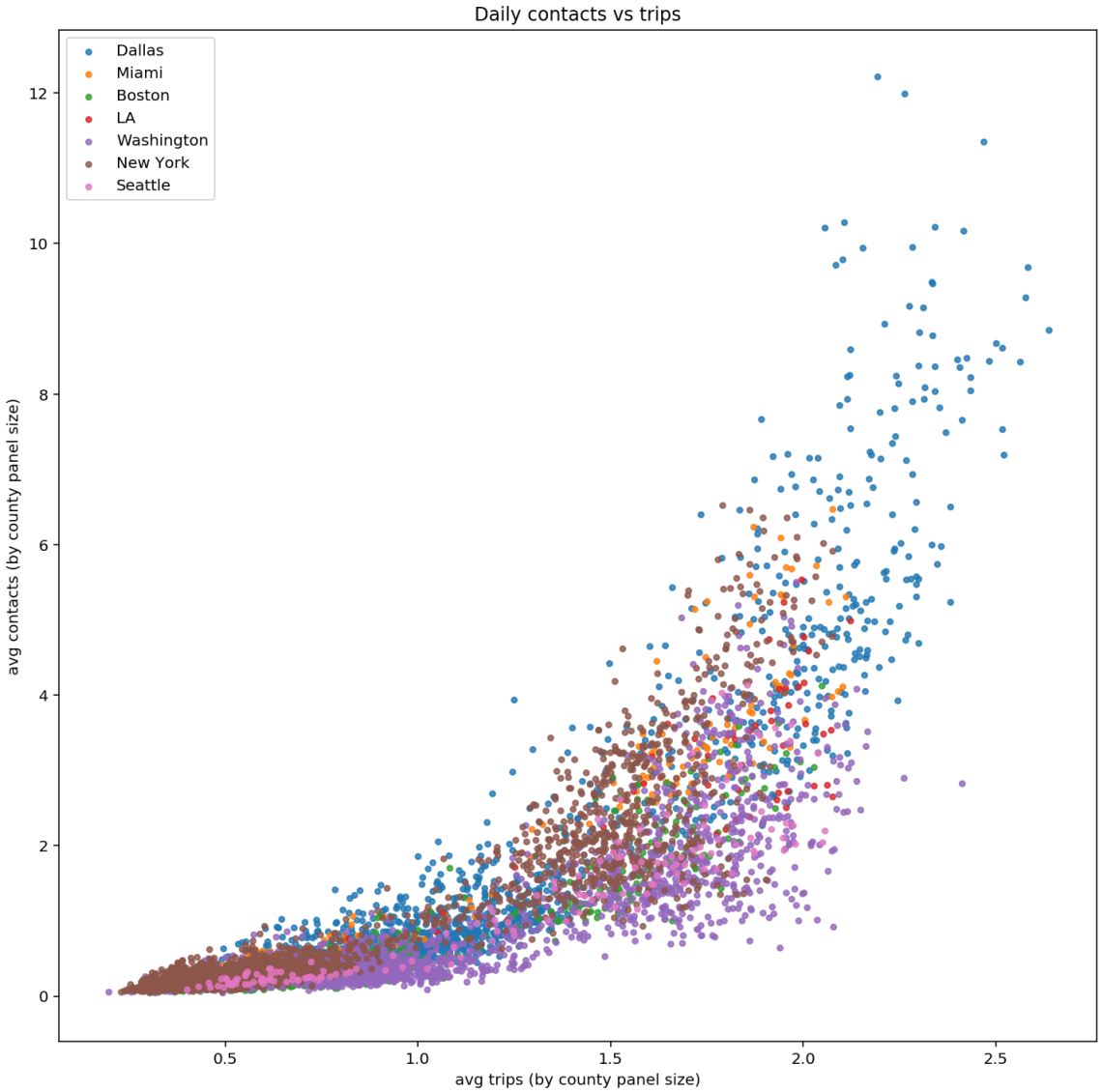


Figure 7-3: Daily contacts and trips metrics shown by plotting daily average trips versus daily average contacts for all counties in each of the 7 metro areas studied. Data points are colored by metro area. Metrics are normalized by the county panel size. We find a relationship between contacts and trips modeled by: $contacts_i(t) \sim trips_i(t)^2$.

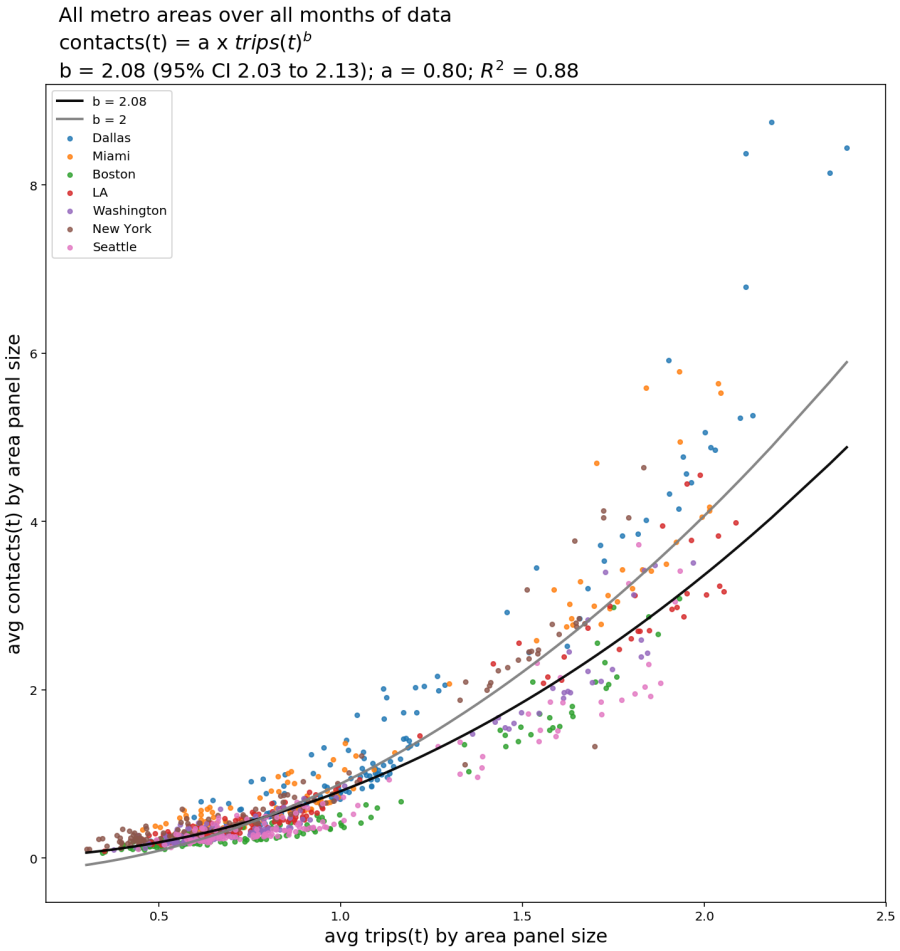


Figure 7-4: Daily contacts vs trips at the metro level, for February 17 through June 27, 2020.

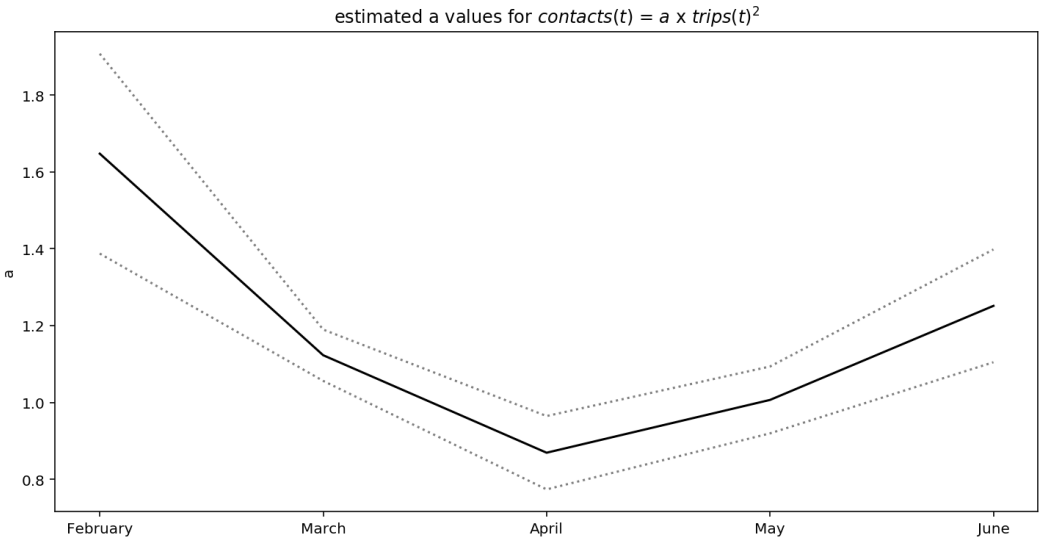


Figure 7-5: In our model, $contacts_i(t) = a \times trips_i(t)^2$, the value of a represents the likelihood of two people making a trip on day t coming into contact. The value of a changes with social distancing behaviors as people make trips in ways to result in fewer contacts (e.g. to less crowded places or spending less time in other places). Our data shows this change occurred. The solid line indicates the estimated value of a, while the dotted lines show the 95% confidence interval.

results are summarized in figure 7-5.

In our model, $contacts_i(t) = a \times trips_i(t)^2$, the value of a represents the likelihood of two people making a trip on day t coming into contact. This value significantly decreased following the social distancing measures that started in March, showing that people made trips in ways to result in fewer contacts. This change reflects changes in social distancing and mobility behaviors consistent with our previous analysis of the New York metro area which showed how the relative portion of trips taken to places by category, such as offices, food destinations, or grocery stores, changed drastically following the national emergency [151]. Our results show that the relationship between contacts and trips continued to change over the months that followed.

Future work can analyze the behavioral changes that drove the change in relationship between the contacts and trips metrics. For example, we can analyze the change in average duration of stays, with the assumption that a shorter duration per trip results in fewer contacts per trip. Consider the case where two people get dinner from the same restaurant. In normal times they would more likely eat there at the same time and result in a "contact".

However, in order to practice social distancing, they might instead both get take-out and stay there for a short time and not come into contact. We can also analyze the change in the spread of time of day people make trips. The new social distancing life has afforded some people more flexibility in when they make trips, such as when they grocery shop, and we can hypothesize that as people make trips to the same places but at more spread out times, this results in fewer contacts per trip. We can also further analyze the changes in the types of places (POIs) visited.

Summary. Overall our results support our theory about the relationship between contacts and trips, which can be estimated by our model $contacts_i(t) = a \times trips_i(t)^2$. However, given the change in the value of a, the contacts metric estimated directly from location based services data is a preferable indicator.

7.4.4 Relationship between contacts and mobility metrics and reported COVID cases

Researchers use mobility metrics such as daily trips and distance traveled to estimate and model the impact of social distancing behavior on growth in COVID-19 cases. For example, time series lead-lag regression analyses are used with the mobility metrics as the independent variable [156, 155, 158] with the expectation that changes in mobility correspond to lagged changes in new cases.

We developed our own lead-lag regression models based on the methods of other researchers. We used each of our metrics computed at the county level - daily trips, distance traveled, and contacts. Our models can be summarized by

$$R_i(t) \sim \beta \times metric_i(t - lag)$$

Where the dependent variable $R_i(t)$ is the estimated growth in new cases on day t for county i, relative to the previous days. Separate models used the daily metrics $trips_i(t)$, $dist_i(t)$, $contacts_i(t)$ to serve as $metric_i$.

However the results of our models and many other researchers' were inconclusive, with results that can be better attributed to fixed effects rather than mobility changes (e.g. [154, 158]).

A simple hypothesis may assume that mobility and reported cases are positively correlated, where decreases in mobility metrics would correlate with relative decreases in growth in reported cases. However, we do not find this to be the case. On the contrary, the data shows the opposite effect. We attribute these issues partly to the limitations and noise of the reported cases data, as well as to the endogenous relationship between reported cases and mobility metrics data. Communities may react to a growth in reported cases with more social distancing and therefore a reduction in contacts and mobility. Additionally, these metrics do not capture other behavioral changes and non-pharmaceutical interventions (NPIs) that complement social distancing, such as the use of masks, increased hygiene, quarantines, reduced international travel, and contact tracing, that can impact the spread of COVID-19 [170, 171, 172].

Figure 7-6 summarizes these findings. New reported COVID-19 cases, contacts, and trips are aggregated by month for each county and relative changes across months are compared[6]. Relative changes are computed as $\frac{(\text{median value for month(i)}) - (\text{median value for month(i - 1)})}{(\text{median value for month(i - 1)})}$ in the figure shown. This is to check if countries that had greater relative reductions in mobility from month to month also had greater relative reductions in new cases month to month. The data does not indicate a positive correlation between these metrics, as a simple hypothesis might expect (correlation is indicated by ρ in the figures). This also indicates why simple lead-lag regression models might not be useful in this kind of analysis.

Despite the limitations for mobility and contacts metrics to serve disease modeling, they may still be useful as monitoring tools. These metrics may help local health organizations preemptively react to potential new cases by monitoring changes in these metrics.

[6]Code notebook that produced these summary figures and results: `https://github.com/aberke/covid-19/blob/master/cases_contacts_mobility_correlations.ipynb`

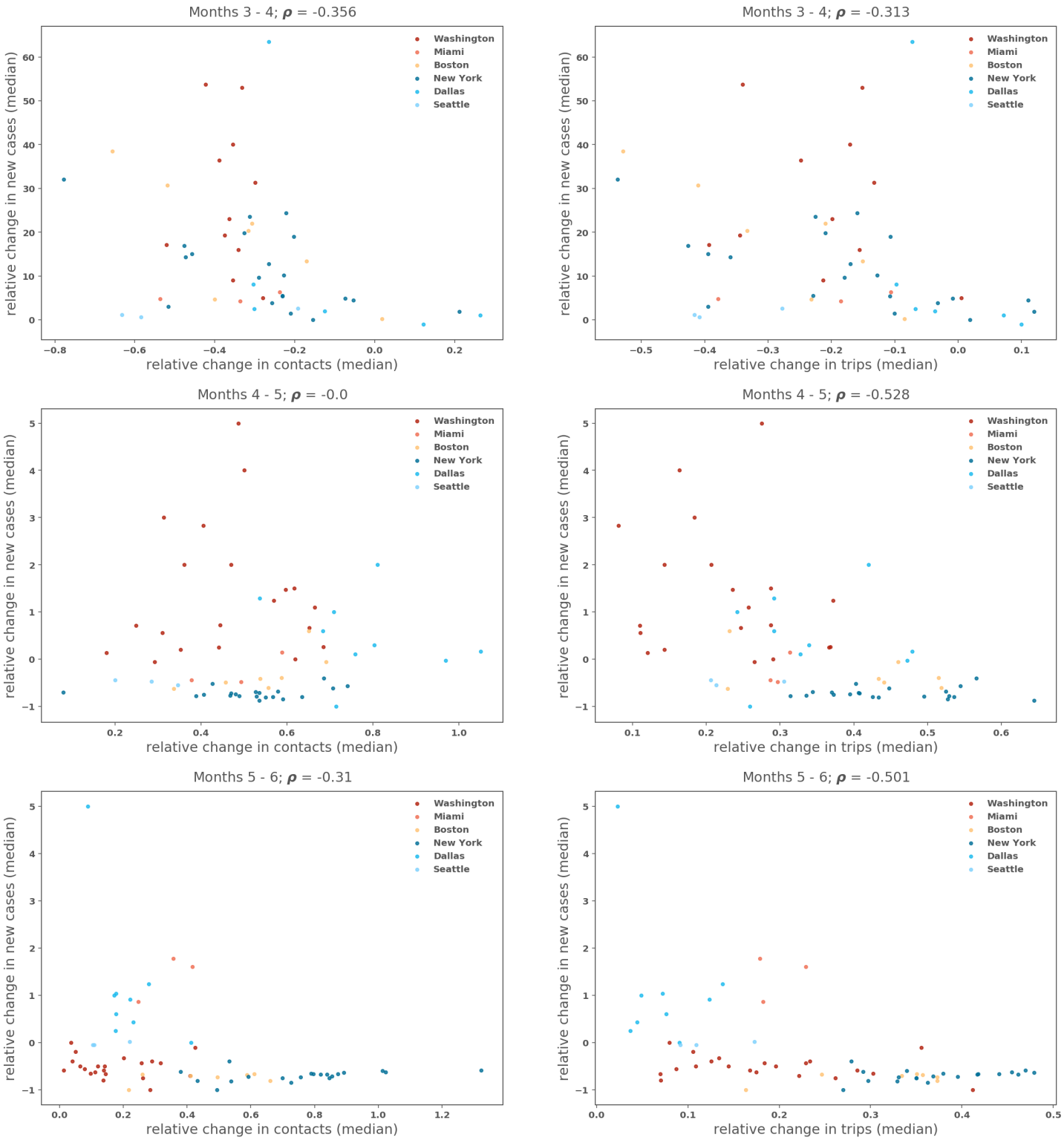


Figure 7-6: Relationship between changes in mobility and reported COVID-19 cases across months: new reported cases, contacts, and trips are aggregated by month for each county and relative changes across months are compared. The correlation (or lack thereof) is indicated by ρ.

7.5 Discussion and Conclusion

In this work we leverage highly granular geolocation data to estimate how often smartphone users are in proximity of one another for an extended duration. We present this "contacts" metric as a more direct proxy for interpersonal contact than the metrics more commonly used by researchers to measure social distancing, as well as to model disease transmission.

We show the relationship between our contacts metric and the more common mobility metrics with a theoretical model which we validate with data from over 1.8 million smartphone users. We also show how this relationship slightly changed over the duration of the crisis. This change in the relationship highlights the importance of more precise metrics, such as the contacts metric, especially when these metrics are used to inform policy decisions.

A limitation of our contacts metric is that it only detects when a limited set of smartphone users come into proximity while allowing their locations to be recorded for a sustained period. As collection and access of geolocation data further increases, the contacts metric can improve.

A limitation of our contacts metric is that it only detects when a limited set of smartphone users come into proximity while allowing their locations to be recorded for a sustained period. As collection and access of geolocation data further increases, the contacts metric can improve. However the high granularity of the geolocation data, which provides for the detection of people coming into contact, also presents privacy risks. For this reason, contacts metrics should only be used or shared in aggregate, such as how this paper aggregates metrics over large geographic areas.

This work was produced in response to the global COVID-19 health pandemic. Researchers around the world are doing important work to understand the impact of events and policies on social distancing behavior and disease transmission. We intend for the contacts metric to aid their work. It can also benefit health agencies, governments, and the people they serve. For example, health agencies may monitor changes in social distancing to preemptively react to potential spikes in new cases. Likewise, better metrics can help governments better understand implications of policy changes and more safely reopen their economies. We hope the contacts metric and work going forward can help stymie the growth of COVID-19 cases

and related disease.

Chapter 8

Conclusion

Large location datasets are being collected from personal devices, primarily by private firms. Should this be the case? How can we design economic incentives and governance models to ethically guide data collection and use? These questions are beyond the scope of this thesis and can be addressed in future work.

Given that these datasets are collected, this thesis addresses how they can serve as public goods. These datasets are collected from the public, and can be used to benefit the public from whom they are sourced in ways similar to surveys collected by government agencies and research organizations. Furthermore, data collected from personal devices can provide unique value that traditional survey data cannot. In particular, this work demonstrates how high-precision location data collected via personal devices can be used to address a public health emergency, as well as serve the public beyond the COVID-19 crisis. Moreover, this work shows how this can be done in ways that protect the privacy of individuals in the datasets.

Will firms that collect location datasets readily democratize their use? Should they?

Again, these questions are beyond the scope of this work. Yet, this thesis presents ways to leverage the unique value these datasets provide without jeopardizing the financial viability of the companies who collect and share the data. For example, chapter 3 shows how real data can be used to generate realistic, privacy-preserving synthetic datasets. The synthetic

data can supplement traditional survey data and provide benefit to the public while the real data remains private. Chapters 6 and 7 demonstrate how location data can be used to better address the impacts of the COVID-19 health crisis. This work leverages the high-precision and real-time nature of location data collected from personal devices to create new metrics that are then used in aggregate. In each case, the companies collecting the data can continue to derive commercial benefit from insights that neither the aggregate metrics nor synthetic datasets can provide. At the same time, privacy can be preserved for real individuals.

Future work. The work in this thesis can be extended to continue to address the COVID-19 health crisis. For example we can use insights from work described in chapters 6 and 7 to further develop methods to better understand the impacts of governmental policy changes and disease transmission. In these chapters we measured interpersonal contacts by leveraging high-precision location data to detect when people came into proximity of one another for a sustained duration. We then showed how the relationship between this contacts metric and other mobility metrics, which were used as proxies for interpersonal contacts by other researchers, changed as the crisis unfolded. This calls into question how well traditional mobility metrics can serve such a crisis as people change their behaviors in order to adapt to it. How can we use data to identify the most relevant metrics for current issues at hand?

In order to accommodate the expanding uses of location data, future work can also expand upon the ways to understand and quantify data privacy. For example, chapter 3 describes how privacy for a location dataset is often quantified by how unique any individual is within the dataset. This is because even when datasets are anonymized, individuals with a unique set of datapoints can be re-identified and private information about their location histories can be revealed. k-anonymity attempts to quantify this privacy risk by how many other individuals within a dataset any individual shares a set of points with. Here a larger group of such individuals means less risk, since any individual within a group cannot be distinguished from any other within the group, and then cannot be re-identified. However, instead of quantifying uniqueness, another perspective considers privacy risks due to the commonalities of such a group. For example, if a group of individuals share sensitive information in their location histories, then distinguishing them from one another is not important. That is, if a group of individuals all went to the same sensitive location (e.g. a protest),

then knowledge that an individual is within the group is enough to expose their sensitive information without the individual being re-identified directly. Alternative privacy frameworks, such as l-diversity, can help address issues such as these. What are better methods to quantify privacy for location datasets? How should these methods address the different *qualitative* risks that different kinds of sensitive locations present? What are ways to share location datasets to address these risks while retaining their utility?

Furthermore, in order for these datasets to serve the public while preserving privacy, we must address the logistics of how they are released. Future work can extend existing privacy-preserving open data publishing models, as well as explore the use of decentralized data ownership models.

As collection of location data from personal devices increases, larger and more comprehensive datasets will be amassed. This will increase the amount of information location datasets can provide, and therefore the amount of utility they can provide the public. At the same time, this will increase privacy risks for individuals and the amount of information that the data can reveal about them. For these reasons, finding new ways to understand and balance the utility and privacy trade-offs for location data will become ever more important in order to benefit the public.

Appendix A

Details in generating synthetic data

A.1 Data representativeness

This work uses data reported by devices in 3 counties (Middlesex, Norfolk, Suffolk counties) surrounding Boston MA. The combined population estimate for these counties is 3,127,354.

The provided data for this geographic area is from 83,827 unique devices, representing about 2.7% of the population. We restrict the data we use to a panel of 22,707 unique user devices that reported at least 3 unique days and 3 unique nights of data during the first 5-day workweek of May 2018. This allows us to more confidently infer home and work locations. The dataset may already have bias, and by selecting for users who report more data, we may introduce new bias to the panel. To better understand the degree of this bias, we do the following. We infer the home census tracts for the users of these devices as the census tract where they spent the most time within between the hours of 8pm and 9am. We compare our resulting census tract population estimates to those provided by the ACS 2018 5-year population estimates (see figure A.1.1). The correlation is ρ=0.648.

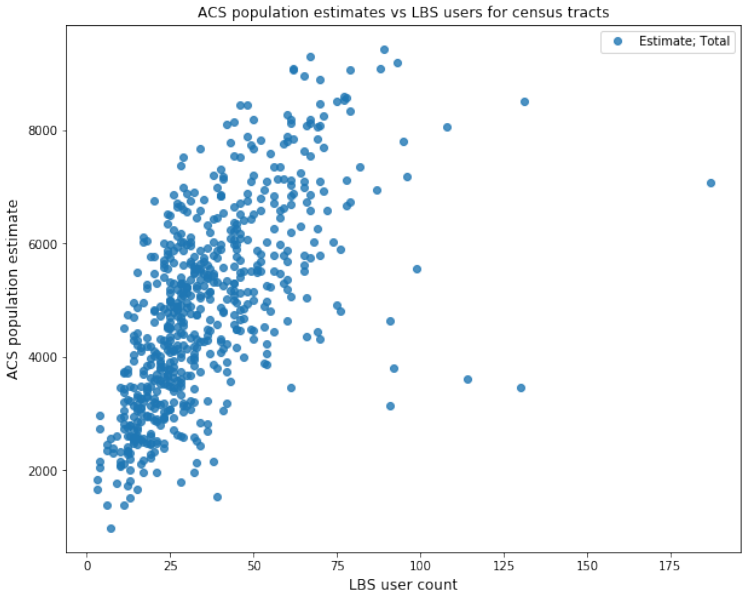


Figure A.1.1: Comparing the numbers of home census tracts that were inferred for users in our data panel to the corresponding census population estimates provided by the American Community Survey 5-Year estimates. The correlation is ρ=0.648.

Appendix B

Technical details for contact tracing technologies

B.1 Privacy-preserving Bluetooth protocols for contact tracing

This appendix section explains the concepts behind the more privacy-preserving Bluetooth protocols designed for contact tracing, such as those from PACT [116], DP-3T [115], and Apple-Google [118]. This section is included to help the reader understand how they work, but also why they are imperfect. These protocols were designed to help keep secret the identities of diagnosed users who share their data, as well as to help protect their movements from being tracked.

These protocols employ similar ideas but differ in their specifics and terminology. This high level explanation avoids the specific differences between protocols and uses terminology that is most consistent with the Apple-Google framework.

At a high level:

Mobile devices broadcast random-looking IDs via Bluetooth Low Energy signals. They also receive IDs broadcast by other devices and record these IDs along with the time at which

they were received. Call these IDs "Rolling Proximity Identifiers", or RPIs for short.

When users are diagnosed as infected, they share information about the RPIs they recently broadcast (from whatever period is determined to be medically relevant) to a "diagnosis server". Other users' apps can then periodically check if any of the RPIs they recently received match the diagnosis server's data. A match indicates that a user came into contact with someone who was later diagnosed and their app can notify them of their exposure risk.

However, diagnosed users do not simply upload their broadcast RPIs to the "diagnosis server". Instead they upload the parameters that generated each RPI.

In computer science and cryptography, one-way and pseudorandom functions (PRFs) are commonly used to hide secrets. Given such a function and its input, it is easy to compute the output. But it is considered computationally infeasible to reverse the function and compute input based on output. A PRF is used to generate the RPIs. Each RPI broadcast by a user's device is the output of a PRF that uses a key (k) that only the device knows and time (t) as inputs.

$$RPI \longleftarrow PRF(k, t)$$

Each RPI is broadcast by a device for only a brief amount of time, after which a new RPI is computed using the latest time as input, and then broadcast instead. Changing the RPI in this way makes it more difficult to track devices or re-identify people who anonymously shared their data to the diagnosis server (but as we'll see, it is still possible).

What the device of a diagnosed user shares to a diagnosis server is the sets of inputs (k, t) that were used to generate the RPIs it broadcast. Another user's device can then use these inputs along with the PRF to recompute the RPIs, and check if any of these RPIs and their corresponding time inputs match against the received RPIs and times that the device stored locally.

You might ask why do it this way? Why not just have users share their RPIs? The answer is to maintain the integrity and security of the system.

Suppose a malicious user wanted to generate false alarms or otherwise create distrust in the system. They might rebroadcast RPIs that were recently uploaded to the diagnosis server. They also might continuously rebroadcast as many RPIs as possible that they received from other users (these malicious behaviors are referred to as "replay attacks"). Other users would then receive the rebroadcast RPIs from the malicious user. If only the RPIs of diagnosed users were uploaded to the diagnosis server rather than their (k, t) inputs, then these users could be falsely notified that they were in contact with diagnosed users. This potential integrity attack is prevented, however, because devices can check that the time they received the RPIs matches against the corresponding (k, t) inputs that the diagnosis server stores.

This protocol also protects users from being framed. Suppose that a user is diagnosed and instead of sharing information for their own RPIs to the diagnosis server, they attempt to dishonestly share RPIs that were broadcast by another user. However, since they must share the (k, t) inputs used to generate the RPIs, this requires knowledge of the other user's key. Since other users can keep their keys secret until they are diagnosed and then choose to share them, this attack is prevented.

The kinds of attacks that the Bluetooth protocol was designed to prevent may seem far-fetched, but keep in mind that the protocol must reliably work without anyone knowing whose data belongs to whom and without depending on a central authority to intervene or prevent misuse. Clever use of cryptography and protocols are then necessary to ensure trust in the system.

This is a very high level overview of how Bluetooth protocols can work to improve the privacy and security of contact tracing systems. There are many more complexities involved in how keys and RPIs are generated. For details of the Apple and Google framework, refer to their specifications: `https://www.apple.com/covid19/contacttracing/`.

For example with the Apple-Google framework, the key used by a single device to produce RPIs periodically changes. This is done so that a user's RPIs will be associated with different sets of keys. Then, when they anonymously share their (k, t) data to the diagnosis server, it will be more difficult to link their data points together, making it more difficult to track them across the locations they visited, and more difficult to re-identify them. However, the protocol is still imperfect and cannot guarantee this type of privacy for its users.

Let's look at another example. If a diagnosed user shares data for an RPI they broadcast that is then received by a contact while receiving no other RPIs, then this contact can easily re-identify them. We can also imagine a future where beacons that listen for Bluetooth signals are sprinkled throughout our environment. (This might be done for a variety of reasons, such as improving contact tracing, or tracking customers in stores and elsewhere to better advertise products.) Users' RPIs could then be recorded throughout the places they go and linked back together once shared, creating a record of their location histories.

Some researchers have proposed using mix networks or private set intersection protocols to mitigate these privacy and security issues [126, 127]. Others have considered reversing the above scheme so that instead of users uploading data for their own broadcast RPIs, they share the RPIs they have received from others (i.e. the "dual approach" [173]). However each of these proposals are imperfect.

Ultimately, tracking people is central to the concept of contact tracing. Many of the newly developed protocols found clever ways to minimize people's loss of privacy while they are tracked, but some privacy may still need to be forfeited for contact tracing to be effective.

B.2 An example private set intersection protocol using Diffie-Hellman

This section outlines an example for how a contact tracing system like the one proposed in chapter 5 could work with a simple Diffie-Hellman private set intersection (PSI) protocol.

Note that the actual implementation could differ from this example. The description assumes familiarity with Diffie-Hellman, modular arithmetic and concepts from cryptography such as the discrete log problem. Otherwise readers can skip to the protocol summary below.

Before we walk through this PSI protocol, we clarify the problem and notation.

Notation and Problem Statement: We call a point interval p, and a collected sequence of point intervals $P = [p_1, p_2, ...]$. We call the users' point intervals that are collected by

their device P_U. We call the point intervals collected for diagnosed carriers and later shared with the server P_I.

As noted earlier, each point interval is encrypted by a commonly shared deterministic hash function, which we call H. This means that a user's phone really stores
$H(P_U) = [H(p_{U1}), H(p_{U1}), ...H(p_{Un})]$, and the server stores
$H(P_I) = [H(p_{I1}), H(p_{I1}), ...H(p_{Im})]$.

If a user has a point interval matching one shared by a diagnosed carrier, i.e. $p_{Ui} = p_{Ij}$ for some p_{Ui} in P_U collected by the user's device, and some p_{Ij} in P_I collected by a diagnosed carriers' device and later shared to the server, then the encrypted hashes of these point intervals match as well. $H(p_{Ui}) = H(p_{Ij})$.

Consider $H(P_U)$ as a set which is stored on the user's device, and $H(P_I)$ as a set stored on the server. The problem is then to allow the user's app to learn the set intersection of $H(P_U)$ and $H(P_I)$.

Protocol: The described use of Diffie-Hellman is written with the multiplicative group of integers modulo p, where p is prime, and g is a primitive root modulo p.

Setup: The server and the client user's device have an agreed upon modulus, p, and base of the multiplicative group, g. The server and client each generate secret private keys, a and b, respectively.

1. The client encrypts the user's hashed point intervals, $H(P_U)$, with a and sends this data to the server.

   $Client \rightarrow Server : H(P_U)^a = [H(p_{U1})^a, H(p_{U1})^a, ..., H(p_{Un})^a] \bmod p$

2. The server encrypts its stored hashed point intervals, $H(P_I)$, with b and sends this data to the client.

   $Server \rightarrow Client : H(P_I)^b = [H(p_{I1})^b, H(p_{I1})^b, ...H(p_{Im})^b] \bmod p$

3. Upon receiving the encrypted hashed point intervals sent by the client $H(P_U)^a$, the server further encrypts this data with its key b and sends the result back to the client.

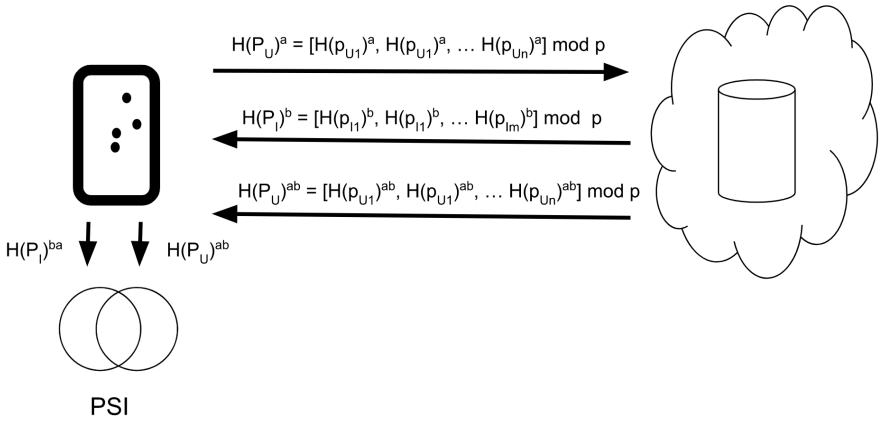


Figure B.2.1: This figure shows our private set intersection (PSI) protocol based on Diffie-Hellman key exchange. The intersection is done on encrypted point intervals corresponding to the client's and carrier's location traces. The encrypted point intervals from diagnosed carriers are stored in a server. Only the client device learns the intersection of data, where the intersection is the set of point intervals it has in common with the point intervals on the server ($P_U \cap P_I$). Points in ($P_U \cap P_I$) are represented in ($H(P_U)^{ab} \cap H(P_I)^{ba}$).

$$Server \rightarrow Client : H(P_U)^{ab} = [H(p_{U1})^{ab}, H(p_{U1})^{ab}, ..., H(p_{Un})^{ab}] \bmod p$$

4. The client receives both $H(P_I)^b$ and $H(P_U)^{ab}$. The client then further encrypts $H(P_I)^b$ with its key a to create

   $H(P_I)^{ba} = [H(p_{I1})^{ba}, H(p_{I1})^{ba}, ..., H(p_{Im})^{ba}] \bmod p.$

5. The client can then compute the set intersection by comparing the elements of $H(P_U)^{ab}$ and $H(P_I)^{ba}$.

Due to the multiplicative properties of the group, any matching $H(p_U)$ and $H(p_I)$ values will have matching $H(p_U)^{ab}$ and $H(p_I)^{ba}$ values. This means that if the client has any point intervals that match point intervals shared to the server, $p_U = p_I$, then $H(p_U)^{ab} = H(p_I)^{ba}$, and these matches will be detected by the client.

Protocol summary

This protocol allows for flexibility in terms of whether it allows a client to learn which of its points have matches versus how many of its points have matches. At step (3) of the protocol the server further encrypts the data received from the client, $H(PU)^a$, and

returns $H(PU)^{ab}$. If the server maintains the order in the sequence of points intervals, then the client can then learn exactly which hashed point intervals in the sequence it sent to the server, $H(P_U)^a = [H(p_{U1})^a, H(p_{U1})^a, ...H(p_{Un})^a]$ match against items in the server's encrypted data, $H(P_I)^{ba}$. If the server instead shuffles the sequence before returning it in step (3), then the client can learn how many of its point intervals match against the server's data, but not which ones do.

In this section I described a simple protocol in order to more easily explain how our proposed system can operate. Optimizations can be made for efficiency. For example, the server can reuse its private key, b, and set of encrypted data across multiple interactions with different clients. It can refresh this key and re-encrypt its data periodically, or as new data is shared by diagnosed carriers or deleted as it becomes old. Decreasing how often the server encrypts its data can increase its efficiency. Faster PSI protocols have been developed, including those that optimize for the exchange of information between a server and client, particularly where the server has a much larger set of data than the client, such as in our use case [174, 141].

Appendix C

Details for the analysis of contacts and mobility metrics

C.1 Additional data details

Data panel and data representativeness[1]

Our analysis included data from 1,826,382 total users from counties surrounding 7 metro areas. Figure C.1.1 shows the scatter plot of the census population (obtained through the ACS 2018 5-year estimates [103]) and the number of users in our panel by county. We observe a Pearson correlation of 0.95[2]. Table C.2 shows information about the metro areas and counties used in our analysis, and the number of users in our data panels.

Reported Cases Data

Figures C.1.2, C.1.3 show reported cases data for each county used in the analysis for the relationship between contacts and mobility metrics and reported cases [169].

[1]Code notebook that produced the figures and tables in this section: `https://github.com/aberke/covid-19/blob/master/cumulative%20cases%20data.ipynb`.

[2]Data for Los Angeles County, California is omitted from the scatter plot due to its high population (10,098,052) but is included in the correlation.

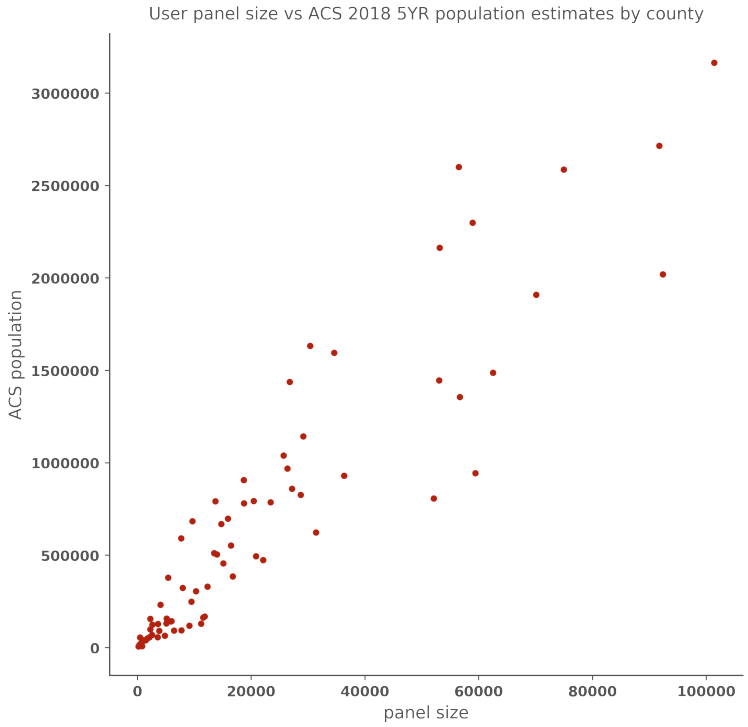


Figure C.1.1: Scatter plot of the number of users in our data panel vs the ACS population estimates for each county. We observe a Pearson correlation of 0.95.

C.2 Contacts and mobility metrics: Additional analysis, details and figures

Distributions for trips and contacts metrics

Figure C.2.4 shows histograms representing the distribution of daily trips and contacts per user. The x-axes show the number of trips and contacts and the y-axes show the number of users in the panel who had that many trips or contacts on the given day. Users who stayed home (i.e. had no trips or contacts) are excluded from the histograms.

The histograms show data for the first weekday and first weekend day for each month that we have data for. This is in order to account for the difference between weekdays and weekends, and show how the distributions slightly changed over time. In general the distributions are highly skewed towards a small number of trips and contacts per user and the distributions have long tails. The distributions become flatter and the tails for the contacts distributions shorten after the month of March, following social distancing measures.

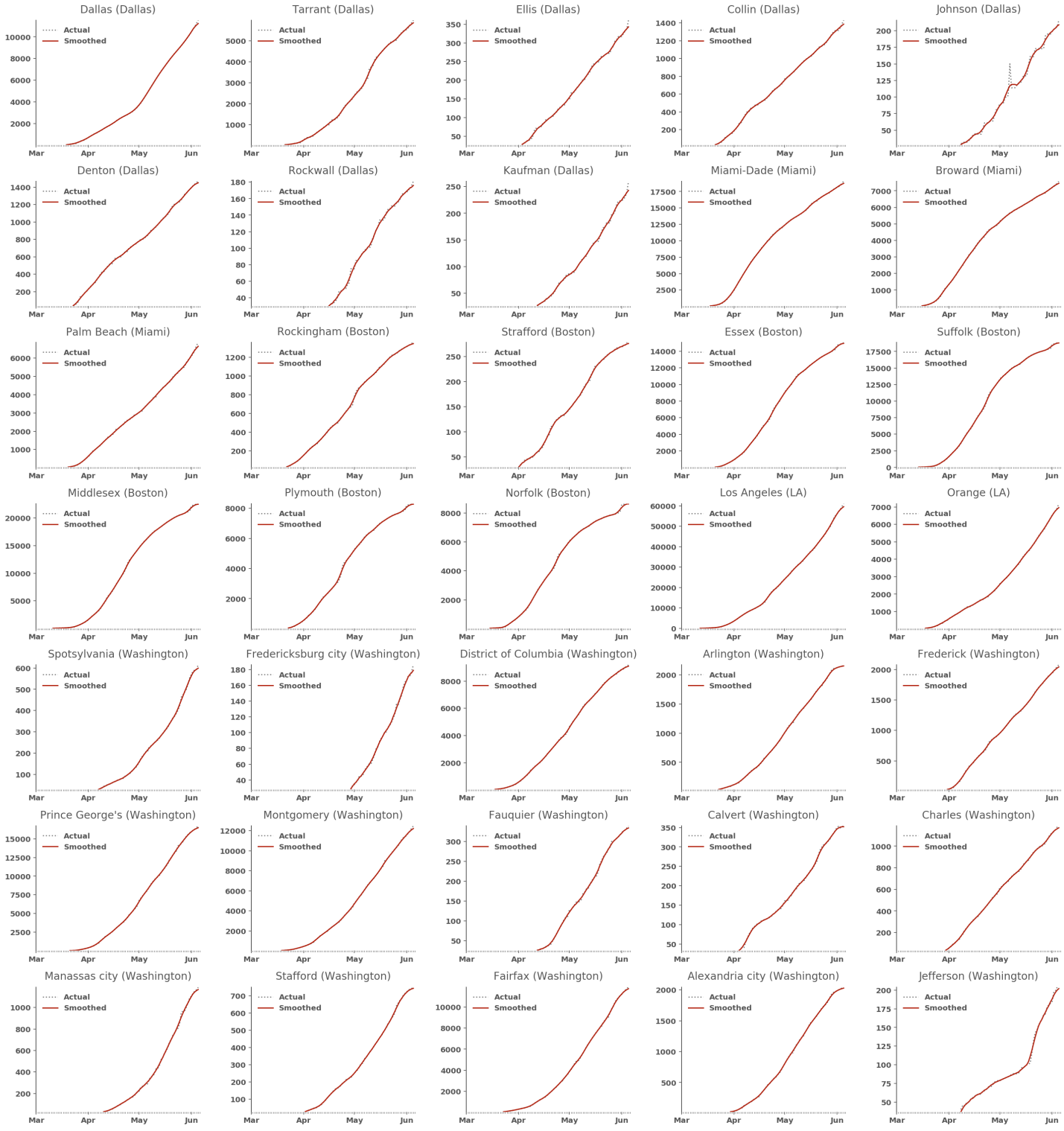


Figure C.1.2: Reported cases data used for counties in analysis.

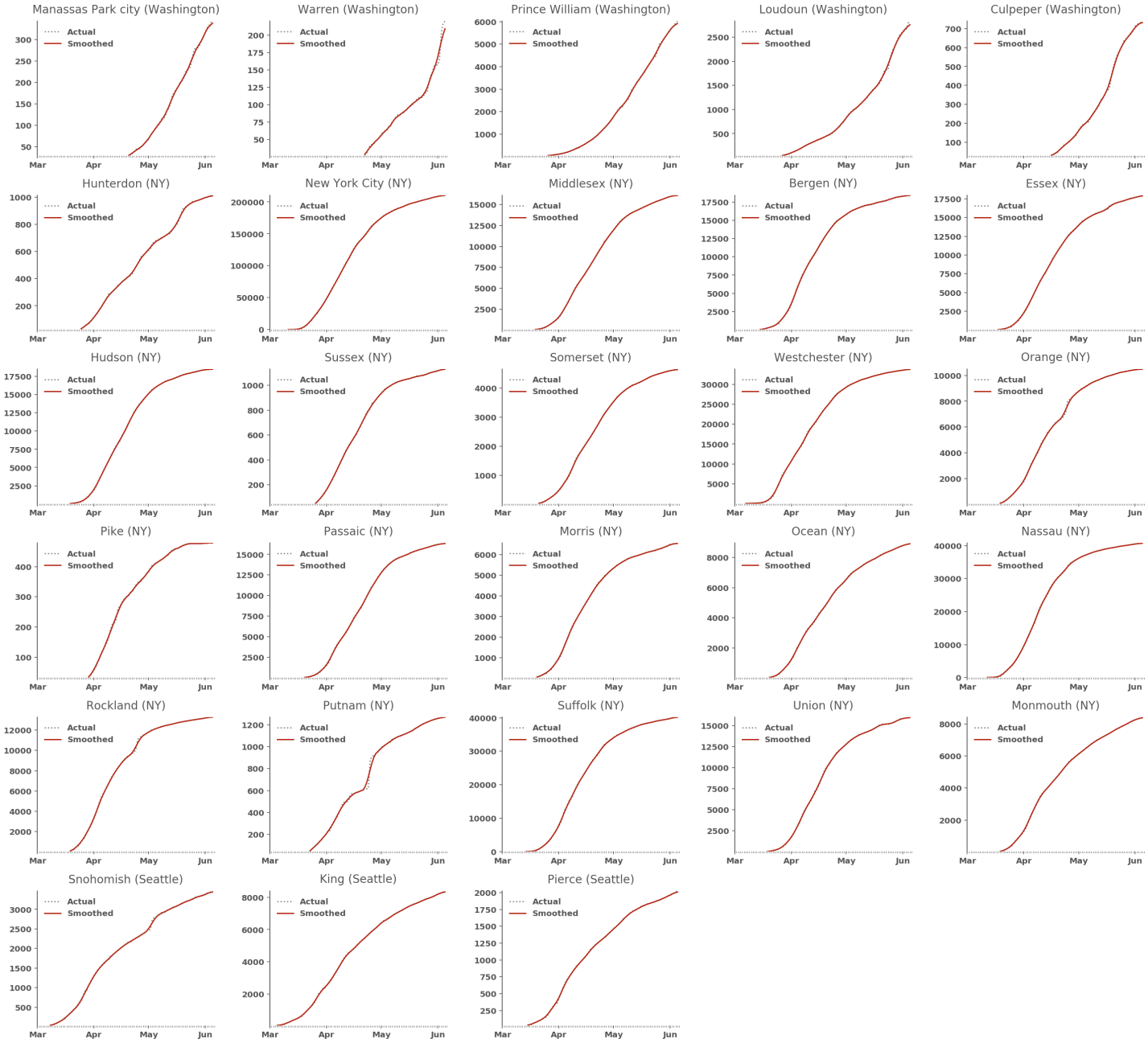


Figure C.1.3: Reported cases data used for counties in analysis (continued).

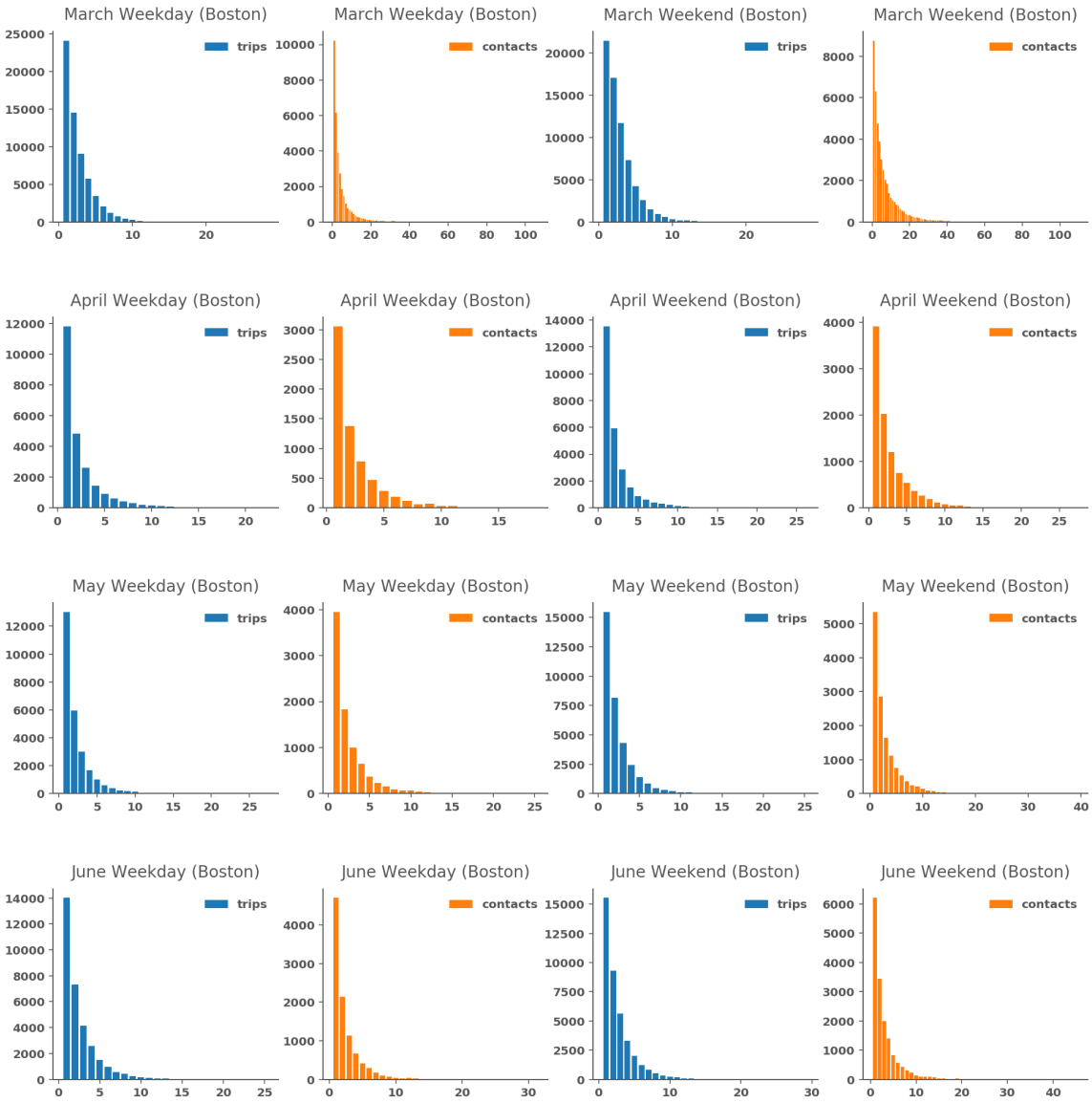


Figure C.2.4: Histograms showing the number of daily trips and contacts per user in the data panel. Data are shown for the first weekday and weekend day for each month we have data for. The number of trips and contacts is on the x-axis. The number of users who reported that many trips or contacts is on the y-axis. Users who stayed home are excluded from the figures.

The plots shown are for the Boston area and are representative of the other metro areas. [3]

Relationship between number of users staying at home and trips [4]

We estimate the number of daily panel users staying at home for each county as the number of total panel users minus the number of distinct panel users with no trips for the given day. We compute metrics at the county level and normalize them by dividing by county panel size. The resulting metrics we use for the following analysis are the portion of daily panel users staying at home, and the average daily trips per user.

Figure C.2.5 shows a plot of daily trips versus users staying at home, for all counties in each metro area.

We hypothesize that there is a linear relationship between the daily number of users *not* staying home and the daily number of trips. To test the hypothesis, we run a regression over the following model to estimate b, with the hypothesis that $b \sim 1$.

$$\text{avg trips} \sim a \times [1 - (\text{portion of users staying home})]^b$$

To do so, we compute the regression as

$$\log(\text{avg trips}) = \log(a) + b \times \log(1 - [\text{portion of users staying home}])$$

Results for the estimated b values for each county are shown in figure C.2.6.

[3] More figures with the distributions of trips and contacts metrics, with more metro areas and more days of data, can be viewed via the code notebook: `https://github.com/aberke/covid-19/blob/master/mobility_contacts_distributions.ipynb`.

[4] Code notebook for this analysis: `https://github.com/aberke/covid-19/blob/master/trips_v_staying_home.ipynb`.

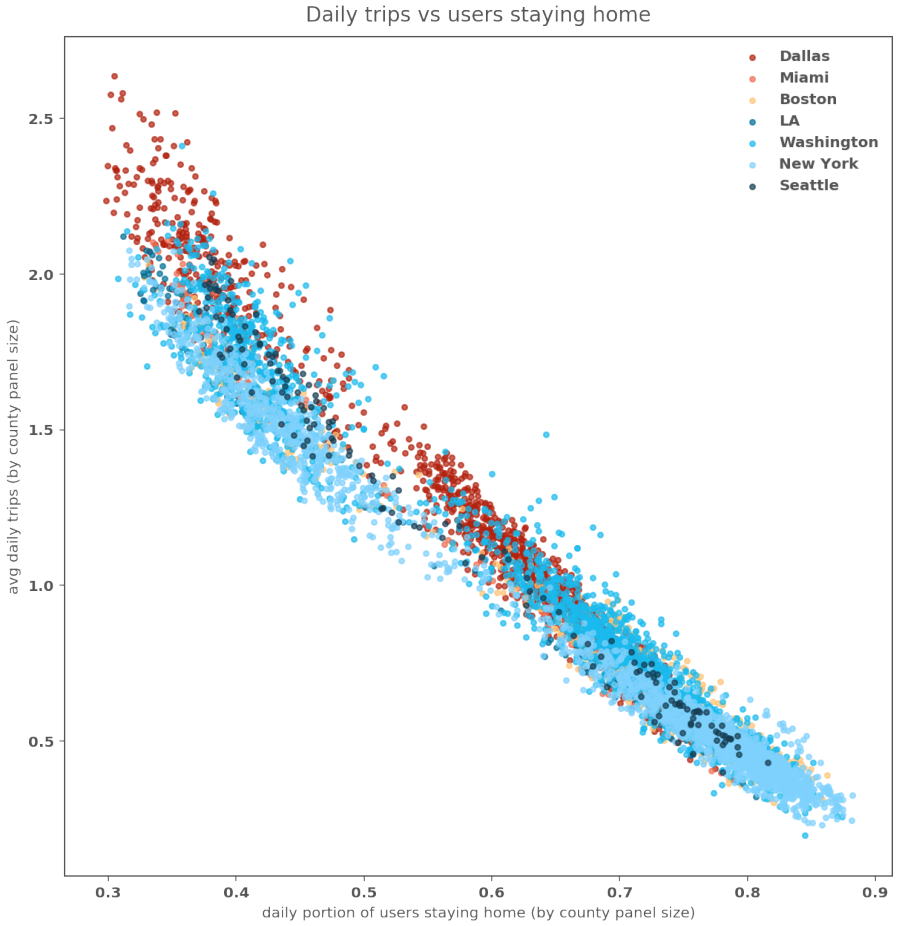


Figure C.2.5: Daily trips versus users staying at home, for all counties in each metro areas.

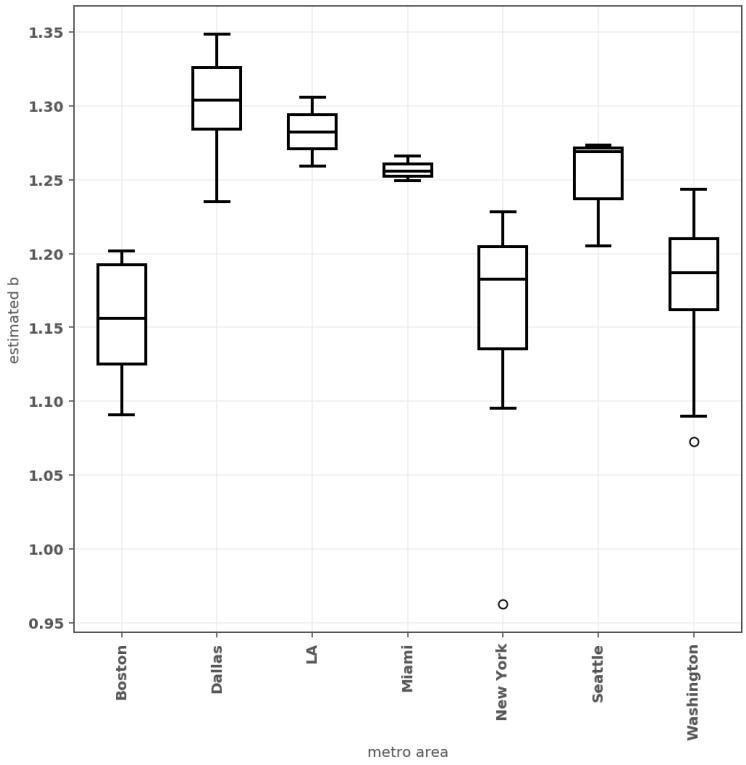


Figure C.2.6: Box plot showing estimated b values for the model $trips \sim a \times [1 - (\text{portion of users staying home})]^b$ computed for each county. Results are grouped by metro area.

	Pooled	**Boston**	**NY**	**Dallas**	**Miami**	**LA**	**DC**	**Seattle**
Dep. variable	$log(trips/dist)$	$log(trips/dist)$	$log(trips/dist)$	$log(trips/dist)$	$log(trips/dist)$	$log(trips/dist)$	$log(trips/dist)$	$log(trips/dist)$
constant	-3.139***	-3.444***	-3.072***	-2.903***	-2.928***	-2.209***	-3.178***	-3.14***
	(0.01)	(0.033)	(0.031)	(0.02)	(0.146)	(0.25)	(0.014)	(0.107)
$log(density)$	0.134***	0.185***	0.128***	0.076***	0.113***	-0.007	0.137***	0.157***
	(0.002)	(0.005)	(0.005)	(0.004)	(0.024)	(0.035)	(0.002)	(0.02)
rsq	0.565	0.7	0.417	0.285	0.072	0.0	0.641	0.267
N	6262	665	1805	884	285	190	2280	153

Table C.1: Results from regression estimating the α value in the model relating the daily trips and distance traveled. $\log(trips_i(t)) - \log(dist_i(t)) = constant + \alpha \times \log(populationDensity_i)$

Distance and Trips Analysis

To estimate the α value for the distance and trips analysis we transform the following equation and use an OLS model. Metrics are computed and used at the county level.

$$trips_i(t) = constant \times populationDensity_i{}^{\alpha} \times dist_i(t)$$

$$\log(trips_i(t)) = constant + \alpha \times \log(populationDensity_i) + \log(dist_i(t))$$

$$\log(trips_i(t)) - \log(dist_i(t)) = constant + \alpha \times \log(populationDensity_i)$$

Results are shown in table C.1.

Table C.2: Panel data used for contacts and mobility data analysis

metro area	county	panel size	population
Dallas	Tarrant County, Texas	92,346	2,019,977
Dallas	Dallas County, Texas	74,937	2,586,552
Dallas	Collin County, Texas	59,413	944,350
Dallas	Denton County, Texas	52,067	807,047
Dallas	Ellis County, Texas	11,797	168,838
Dallas	Johnson County, Texas	11,556	163,475
Dallas	Parker County, Texas	11,166	129,802
Dallas	Kaufman County, Texas	9,104	118,910
Dallas	Rockwall County, Texas	7,740	93,642
Dallas	Hunt County, Texas	6,428	92,152
Dallas	Wise County, Texas	4,816	64,639
Dallas	Hood County, Texas	3,530	56,901
Dallas	Somervell County, Texas	811	8,743
Seattle	King County, Washington	53,136	2,163,257
Seattle	Pierce County, Washington	27,187	859,840
Seattle	Snohomish County, Washington	23,413	786,620
Washington	Fairfax County, Virginia	29,146	1,143,529
Washington	Montgomery County, Maryland	25,676	1,040,133
Washington	Prince George's County, Maryland	18,673	906,202
Washington	Loudoun County, Virginia	16,732	385,143
Washington	Prince William County, Virginia	15,108	456,749
Washington	District of Columbia, District of Columbia	9,656	684,498
Washington	Frederick County, Maryland	9,480	248,472
Washington	Stafford County, Virginia	5,975	144,012
Washington	Charles County, Maryland	5,117	157,671
Washington	Spotsylvania County, Virginia	5,074	131,412
Washington	Arlington County, Virginia	4,064	231,803
Washington	Calvert County, Maryland	3,818	91,082
Washington	Fauquier County, Virginia	2,542	69,115
Washington	Alexandria city, Virginia	2,229	156,505
Washington	Jefferson County, West Virginia	2,135	56,179
Washington	Culpeper County, Virginia	1,701	50,450
Washington	Warren County, Virginia	1,429	39,449
Washington	Manassas city, Virginia	881	41,457
Washington	Fredericksburg city, Virginia	773	28,469
Washington	Fairfax city, Virginia	615	23,865
Washington	Clarke County, Virginia	492	14,365
Washington	Manassas Park city, Virginia	355	16,423
Washington	Falls Church city, Virginia	309	14,067
Washington	Rappahannock County, Virginia	165	7,332
New York City	Suffolk County, New York	62,488	1,487,901
New York City	Queens County, New York	58,916	2,298,513

New York City	Nassau County, New York	56,660	1,356,564
New York City	Kings County, New York	56,499	2,600,747
New York City	Bergen County, New Jersey	36,303	929,999
New York City	Monmouth County, New Jersey	31,367	623,387
New York City	New York County, New York	30,349	1,632,480
New York City	Middlesex County, New Jersey	28,695	826,698
New York City	Bronx County, New York	26,767	1,437,872
New York City	Westchester County, New York	26,362	968,815
New York City	Richmond County, New York	22,071	474,101
New York City	Morris County, New Jersey	20,820	494,383
New York City	Essex County, New Jersey	20,434	793,555
New York City	Union County, New Jersey	16,448	553,066
New York City	Hudson County, New Jersey	14,721	668,631
New York City	Passaic County, New Jersey	13,961	504,041
New York City	Somerset County, New Jersey	12,323	330,176
New York City	Rockland County, New York	7,947	323,686
New York City	Ocean County, New Jersey	7,667	591,939
New York City	Sussex County, New Jersey	5,872	142,298
New York City	Orange County, New York	5,413	378,227
New York City	Hunterdon County, New Jersey	2,594	125,051
New York City	Putnam County, New York	2,242	99,070
New York City	Pike County, Pennsylvania	470	55,498
Miami	Miami-Dade County, Florida	91,736	2,715,516
Miami	Broward County, Florida	70,080	1,909,151
Miami	Palm Beach County, Florida	53,017	1,446,277
Boston	Middlesex County, Massachusetts	34,595	1,595,192
Boston	Essex County, Massachusetts	18,721	781,024
Boston	Norfolk County, Massachusetts	15,889	698,249
Boston	Suffolk County, Massachusetts	13,707	791,766
Boston	Plymouth County, Massachusetts	13,456	512,135
Boston	Rockingham County, New Hampshire	10,301	305,129
Boston	Strafford County, New Hampshire	3,615	128,237
LA	Los Angeles County, California	220,900	10,098,052
LA	Orange County, California	101,384	3,164,182

Bibliography

[1] United States Census Bureau. About the bureau: What we do. https://www.census.gov/about/what.html. Accessed: July 2020.

[2] Sen Pei, Sasikiran Kandula, and Jeffrey Shaman. Differential effects of intervention timing on covid-19 spread in the united states. *medRxiv*, 2020.

[3] Bluebikes system data. https://www.bluebikes.com/system-data. Accessed: July 2020.

[4] Nyc taxi and limousine commission: Tlc trip record data. https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Accessed: July 2020.

[5] New York State. Turnstile usage data for brooklyn-manhattan transit (bmt) division: 2016. https://data.ny.gov/Transportation/Turnstile-Usage-Data-for-Brooklyn-Manhattan-Transi/ev32-4fwz. Accessed: July 2020.

[6] Transport For London. Open data policy. https://tfl.gov.uk/info-for/open-data-users/open-data-policy. Accessed: July 2020.

[7] LLP Deloitte. Assessing the value of tfl's open data and digital partnerships, 2017.

[8] Andrew J Blumberg and Peter Eckersley. On locational privacy, and how to avoid losing it forever. *Electronic frontier foundation*, 10(11):1–7, 2009.

[9] Marco Fiore, Panagiota Katsikouli, Elli Zavou, Mathieu Cunche, Françoise Fessant, Dominique Le Hello, Ulrich Matchi Aivodji, Baptiste Olivier, Tony Quertier, and Razvan Stanica. Privacy in trajectory micro-data publishing: a survey. *arXiv preprint arXiv:1903.12211*, 2019.

[10] United States Census Bureau. Decennial census of population and housing. https://www.census.gov/programs-surveys/decennial-census.html. Accessed: July 2020.

[11] United States Census Bureau. About the american community survey. https://www.census.gov/programs-surveys/acs/about.html. Accessed: July 2020.

[12] Paul Boyle and Danny Dorling. Guest editorial: the 2001 uk census: remarkable resource or bygone legacy of the 'pencil and paper era'? *Area*, 36(2):101–110, 2004.

[13] David Coleman. The twilight of the census. *Population and development review*, 38:334–351, 2013.

[14] Robert M Groves. Nonresponse rates and nonresponse bias in household surveys. *Public opinion quarterly*, 70(5):646–675, 2006.

[15] David Cantor, Gary Shapiro, L Chen, G Choudhry, and Mark Freedman. Non-response in the national household transportation survey (nhts). In *Transportation Research Board. National Household Travel Survey Conference, Washington, DC, November*, pages 1–2, 2004.

[16] Nuria Oliver, Aleksandar Matic, and Enrique Frias-Martinez. Mobile network data for public health: opportunities and challenges. *Frontiers in public health*, 3:189, 2015.

[17] United Kingdom Department of Transportation. National travel survey quality report. `https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/775062/annex-d-nts-2019-quality-report.pdf`.

[18] Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R Stevens, Andrea E Gaughan, Vincent D Blondel, and Andrew J Tatem. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45):15888–15893, 2014.

[19] United Kingdom Office for National Statistics. Ons methodology working paper series no. 8 - statistical uses for mobile phone data: literature review. `https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries`, 2016. Accessed: July 2020.

[20] UN Global Working Group on Big Data for Official Statistics. Handbook on the use of mobile phone data for official statistics (draft). `https://unstats.un.org/bigdata/taskteams/mobilephone/MPD%20Handbook%2020191004.pdf`, 2019. Accessed: July 2020.

[21] AirSage. The future of transportation studies: A comparative review. 2013.

[22] Serdar Çolak, Lauren P Alexander, Bernardo G Alvim, Shomik R Mehndiratta, and Marta C González. Analyzing cell phone location data for urban travel: current methods, limitations, and opportunities. *Transportation Research Record*, 2526(1):126–135, 2015.

[23] United States Federal Highway Administration. National household travel survey. `https://nhts.ornl.gov/`. Accessed: July 2020.

[24] Westat. 2017 nhts data user guide. `https://nhts.ornl.gov/assets/2017UsersGuide.pdf`. Accessed: July 2020.

[25] Westat. 2017 nhts weighting report. `https://nhts.ornl.gov/assets/2017%20NHTS%20Weighting%20Report.pdf`. Accessed: July 2020.

[26] United States Federal Highway Administration. National household travel survey compendium of uses (2017). `https://nhts.ornl.gov/2017/pub/Compendium_2017.pdf`. Accessed: July 2020.

[27] Edward Barbour, Carlos Cerezo Davila, Siddharth Gupta, Christoph Reinhart, Jasleen Kaur, and Marta C González. Planning for sustainable cities by estimating building occupancy with mobile phones. *Nature communications*, 10(1):1–10, 2019.

[28] Lauren Alexander, Shan Jiang, Mikel Murga, and Marta C González. Origin-destination trips by purpose and time of day inferred from mobile phone data. *Transportation research part c: emerging technologies*, 58:240–250, 2015.

[29] GroundTruth. How puredriven leaned into location-based marketing to strategically drive in-store visits. `http://go.groundtruth.com/puredriven`. Accessed: July 2020.

[30] Dan Calacci, Alex Berke, Kent Larson, et al. The tradeoff between the utility and risk of location data and implications for public good. *arXiv preprint arXiv:1905.09350*, 2019.

[31] Claudio Bettini, X Sean Wang, and Sushil Jajodia. Protecting privacy against location-based personal identification. In *Workshop on Secure Data Management*, pages 185–199. Springer, 2005.

[32] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3:1376, 2013.

[33] Anthony Tockar. Riding with the stars: Passenger privacy in the nyc taxicab dataset. *Neustar Research, September*, 15, 2014.

[34] Mudhakar Srivatsa and Mike Hicks. Deanonymizing mobility traces: Using social network as a side-channel. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pages 628–637, 2012.

[35] John Krumm. Inference attacks on location tracks. In *International Conference on Pervasive Computing*, pages 127–143. Springer, 2007.

[36] Baik Hoh, Marco Gruteser, Hui Xiong, and Ansaf Alrabady. Enhancing security and privacy in traffic-monitoring systems. *IEEE Pervasive Computing*, 5(4):38–46, 2006.

[37] Chris YT Ma, David KY Yau, Nung Kwan Yip, and Nageswara SV Rao. Privacy vulnerability of published anonymous mobility traces. In *Proceedings of the sixteenth annual international conference on Mobile computing and networking*, pages 185–196, 2010.

[38] Luca Rossi, James Walker, and Mirco Musolesi. Spatio-temporal techniques for user identification by means of gps mobility data. *EPJ Data Science*, 4(1):1–16, 2015.

[39] Takao Murakami, Atsunori Kanemura, and Hideitsu Hino. Group sparsity tensor factorization for re-identification of open mobility traces. *IEEE Transactions on Information Forensics and Security*, 12(3):689–704, 2016.

[40] Yi Song, Daniel Dahlmeier, and Stephane Bressan. Not so unique in the crowd: a simple and effective algorithm for anonymizing location data. *PIR@ SIGIR*, 2014:19–24, 2014.

[41] Julián Salas, David Megías, and Vicenç Torra. Swapmob: Swapping trajectories for mobility anonymization. In *International Conference on Privacy in Statistical Databases*, pages 331–346. Springer, 2018.

[42] Md Shahadat Iqbal, Charisma F Choudhury, Pu Wang, and Marta C González. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40:63–74, 2014.

[43] Markus Friedrich, Katrin Immisch, Prokop Jehlicka, Thomas Otterstätter, and Johannes Schlaich. Generating origin–destination matrices from mobile phone trajectories. *Transportation research record*, 2196(1):93–101, 2010.

[44] Tian-ran Hu, Jie-bo Luo, Henry Kautz, and Adam Sadilek. Home location inference from sparse and noisy data: models and applications. *Frontiers of Information Technology & Electronic Engineering*, 17(5):389–402, 2016.

[45] Kevin S Kung, Kael Greco, Stanislav Sobolevsky, and Carlo Ratti. Exploring universal patterns in human home-work commuting from mobile phone data. *PloS one*, 9(6):e96180, 2014.

[46] Shan Jiang, Yingxiang Yang, Siddharth Gupta, Daniele Veneziano, Shounak Athavale, and Marta C González. The timegeo modeling framework for urban mobility without travel surveys. *Proceedings of the National Academy of Sciences*, 113(37):E5370–E5378, 2016.

[47] Luca Pappalardo and Filippo Simini. Data-driven generation of spatio-temporal routines in human mobility. *Data Mining and Knowledge Discovery*, 32(3):787–829, 2018.

[48] Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, 2010.

[49] Hugo Barbosa, Fernando B de Lima-Neto, Alexandre Evsukoff, and Ronaldo Menezes. The effect of recency to human mobility. *EPJ Data Science*, 4:1–14, 2015.

[50] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *nature*, 453(7196):779–782, 2008.

[51] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[52] Rui Chen, Benjamin CM Fung, Bipin C Desai, and Nériah M Sossou. Differentially private transit data publication: a case study on the montreal transportation system. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 213–221, 2012.

[53] Rui Chen, Gergely Acs, and Claude Castelluccia. Differentially private sequential data publication via variable-length n-grams. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pages 638–649, 2012.

[54] Mehmet Emre Gursoy, Ling Liu, Stacey Truex, and Lei Yu. Differentially private and utility preserving publication of trajectory data. *IEEE Transactions on Mobile Computing*, 18(10):2315–2329, 2018.

[55] Xi He, Graham Cormode, Ashwin Machanavajjhala, Cecilia M Procopiuc, and Divesh Srivastava. Dpt: differentially private trajectory synthesis using hierarchical reference systems. *Proceedings of the VLDB Endowment*, 8(11):1154–1165, 2015.

[56] Darakhshan J Mir, Sibren Isaacman, Ramón Cáceres, Margaret Martonosi, and Rebecca N Wright. Dp-where: Differentially private modeling of human mobility. In *2013 IEEE international conference on big data*, pages 580–588. IEEE, 2013.

[57] Nazmiye Ceren Abay, Yan Zhou, Murat Kantarcioglu, Bhavani Thuraisingham, and Latanya Sweeney. Privacy preserving synthetic data release using deep learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 510–526. Springer, 2018.

[58] Jun Zhang, Graham Cormode, Cecilia M Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes: Private data release via bayesian networks. *ACM Transactions on Database Systems (TODS)*, 42(4):1–41, 2017.

[59] Gergely Acs, Luca Melis, Claude Castelluccia, and Emiliano De Cristofaro. Differentially private mixture of generative neural networks. *IEEE Transactions on Knowledge and Data Engineering*, 31(6):1109–1121, 2018.

[60] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.

[61] Reihaneh Torkzadehmahani, Peter Kairouz, and Benedict Paten. Dp-cgan: Differentially private synthetic data and label generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[62] Haoran Li, Li Xiong, Lifan Zhang, and Xiaoqian Jiang. Dpsynthesizer: differentially private data synthesizer for privacy preserving data sharing. In *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, volume 7, page 1677. NIH Public Access, 2014.

[63] Abhinav Jauhri, Brad Stocks, Jian Hui Li, Koichi Yamada, and John Paul Shen. Using gans for generation of realistic city-scale ride sharing/hailing data sets. 2018.

[64] Jiawei Wang, Ruixiang Chen, and Zhaocheng He. Traffic speed prediction for urban transportation network: A path based deep learning approach. *Transportation Research Part C: Emerging Technologies*, 100:372–385, 2019.

[65] Qixiu Cheng, Yang Liu, Wei Wei, and Zhiyuan Liu. Analysis and forecasting of the day-to-day travel demand variations for large-scale transportation networks: a deep learning approach. *Transportation Analytics Contest, Tech. Rep*, 2016.

[66] Jinsung Yoon, Daniel Jarrett, and Mihaela van der Schaar. Time-series generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 5508–5518, 2019.

[67] Cristóbal Esteban, Stephanie L Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633*, 2017.

[68] Vaibhav Kulkarni and Benoît Garbinato. Generating synthetic mobility traffic using rnns. In *Proceedings of the 1st Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery*, pages 1–4, 2017.

[69] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[70] Christian M Schneider, Vitaly Belik, Thomas Couronné, Zbigniew Smoreda, and Marta C González. Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, 10(84):20130246, 2013.

[71] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[72] Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

[73] Andrej Karpathy. The unreasonable effectiveness of recurrent neural networks. `https://karpathy.github.io/2015/05/21/rnn-effectiveness/`.

[74] Ilya Sutskever, James Martens, and Geoffrey E Hinton. Generating text with recurrent neural networks. In *ICML*, 2011.

[75] Nipun Agarwala, Yuki Inoue, and Alex Sly. Music composition using recurrent neural networks. *CS 224n: Natural Language Processing with Deep Learning, Spring*, 2017.

[76] Douglas Eck and Juergen Schmidhuber. A first look at music composition using lstm recurrent neural networks. *Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale*, 103:48, 2002.

[77] Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. *arXiv preprint arXiv:1206.6392*, 2012.

[78] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[79] Vincent Bindschaedler, Reza Shokri, and Carl A Gunter. Plausible deniability for privacy-preserving data synthesis. *arXiv preprint arXiv:1708.07975*, 2017.

[80] Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

[81] Christof Ferreira Torres and Rolando Trujillo-Rasua. The fréchet/manhattan distance and the trajectory anonymisation problem. In *IFIP Annual Conference on Data and Applications Security and Privacy*, pages 19–34. Springer, 2016.

[82] Xiaokui Xiao, Guozhang Wang, and Johannes Gehrke. Differential privacy via wavelet transforms. *IEEE Transactions on knowledge and data engineering*, 23(8):1200–1214, 2010.

[83] Thomas Brinkhoff. A framework for generating network-based moving objects. *GeoInformatica*, 6(2):153–180, 2002.

[84] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[85] Manolis Terrovitis and Nikos Mamoulis. Privacy preservation in the publication of trajectories. In *The Ninth International Conference on Mobile Data Management (mdm 2008)*, pages 65–72. IEEE, 2008.

[86] Marco Gramaglia and Marco Fiore. Hiding mobile traffic fingerprints with glove. In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies*, pages 1–13, 2015.

[87] Anna Monreale, Gennady L Andrienko, Natalia V Andrienko, Fosca Giannotti, Dino Pedreschi, Salvatore Rinzivillo, and Stefan Wrobel. Movement data anonymity through generalization. *Trans. Data Priv.*, 3(2):91–121, 2010.

[88] Josep Domingo-Ferrer and Rolando Trujillo-Rasua. Microaggregation-and permutation-based anonymization of movement data. *Information Sciences*, 208:55–80, 2012.

[89] Apple. Apple differential privacy technical overview. `https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf`. Accessed July, 2020.

[90] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014.

[91] John M Abowd. Protecting the confidentiality of america's statistics: Adopting modern disclosure avoidance methods at the census bureau. *Census Blogs: Research Matters*, 2018.

[92] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

[93] Mijung Park, James Foulds, Kamalika Choudhary, and Max Welling. Dp-em: Differentially private expectation maximization. In *Artificial Intelligence and Statistics*, pages 896–904, 2017.

[94] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.

[95] Vincent Bindschaedler and Reza Shokri. Synthesizing plausible privacy-preserving location traces. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 546–563. IEEE, 2016.

[96] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

[97] Zhouyu Fu, Weiming Hu, and Tieniu Tan. Similarity based vehicle trajectory clustering and anomaly detection. In *IEEE International Conference on Image Processing 2005*, volume 2, pages II–602. Ieee, 2005.

[98] Dan Buzan, Stan Sclaroff, and George Kollios. Extraction and clustering of motion trajectories in video. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 2, pages 521–524. IEEE, 2004.

[99] Imran N Junejo, Omar Javed, and Mubarak Shah. Multi feature path modeling for video surveillance. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 2, pages 716–719. IEEE, 2004.

[100] Jiangung Lou, Qifeng Liu, Tieniu Tan, and Weiming Hu. Semantic interpretation of object activities in a surveillance system. In *Object recognition supported by user interaction for service robots*, volume 3, pages 777–780. IEEE, 2002.

[101] Zhang Zhang, Kaiqi Huang, and Tieniu Tan. Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 1135–1138. IEEE, 2006.

[102] Lei Chen, M Tamer Özsu, and Vincent Oria. Robust and fast similarity search for moving object trajectories. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 491–502, 2005.

[103] U.S. Census Bureau. American community survey 2014-2018 5-year estimates. `https://data.census.gov/cedsci/table`.

[104] Philippe Golle and Kurt Partridge. On the anonymity of home/work location pairs. In *International Conference on Pervasive Computing*, pages 390–397. Springer, 2009.

[105] Feilong Wang, Jingxing Wang, Jinzhou Cao, Cynthia Chen, and Xuegang Jeff Ban. Extracting trips from multi-sourced data for mobility pattern analysis: An app-based data example. *Transportation Research Part C: Emerging Technologies*, 105:183–202, 2019.

[106] Alex Berke and Kent Larson. Contact tracing technologies: Methods and trade-offs. April 2020.

[107] Young Joon Park, Young June Choe, Ok Park, Shin Young Park, Young-Man Kim, Jieun Kim, Sanghui Kweon, Yeonhee Woo, Jin Gwack, Seong Sun Kim, et al. Contact tracing during coronavirus disease outbreak, south korea, 2020. *Emerging infectious diseases*, 26(10), 2020.

[108] Forbes. South korea's widespread testing and contact tracing lead to first day with no new cases. `https://www.forbes.com/sites/alexandrasternlicht/2020/04/30/south-koreas-widespread-testing-and-contact-tracing-lead-to-first-day-with-no-new-cases/#642283a75abf`, 2020. Accessed: March 2020.

[109] M Zastrow. South korea is reporting intimate details of covid-19 cases: has it helped? *Nature*, 2020.

[110] The New Yorker. Seoul's radical experiment in digital contact tracing. `https://www.newyorker.com/news/news-desk/seouls-radical-experiment-in-digital-contact-tracing`, 2020. Accessed: March 2020.

[111] The New York Times. In coronavirus fight, china gives citizens a color code, with red flags. `https://www.nytimes.com/2020/03/01/business/china-coronavirus-surveillance.html`, 2020. Accessed: March 2020.

[112] Government Digital Services and Blue Trace. Tracetogether. `https://www.tracetogether.gov.sg/`, 2020. Accessed: March 2020.

[113] Dana M. Lewis et al. Coepi: Community epidemiology in action. https://github.com/Co-Epi, 2020. Accessed: March 2020.

[114] Rhys Fenwick, Mike Hittle, Mark Ingle, Oliver Nash, Victoria Nguyen, James Petrie, Jeff Schwaber, Zsombor Szabo, Akhil Veeraghanta, Mikhail Voloshin, Sydney Von Arx, and Tina White. Covid watch. https://www.covid-watch.org/, 2020. Accessed: March 2020.

[115] Carmela Troncoso, Mathias Payer, Jean-Pierre Hubaux, Marcel Salathé, James Larus, Edouard Bugnion, Wouter Lueks, Theresa Stadler, Apostolos Pyrgelis, Daniele Antonioli, et al. Decentralized privacy-preserving proximity tracing. *arXiv preprint arXiv:2005.12273*, 2020.

[116] Ronald L Rivest, Jon Callas, Ran Canetti, Kevin Esvelt, Daniel Kahn Gillmor, Yael Tauman Kalai, Anna Lysyanskaya, Adam Norige, Ramesh Raskar, Adi Shamir, et al. The pact protocol specification, 2020.

[117] Scott Leibrand Jack Gallagher Hamish Manu Eder Zsombor Szabo George Danezis (UCL) Ian Miers Henry de Valence Daniel Reusche Sourabh Niyogi, James Petrie. Tcn protocol. https://github.com/TCNCoalition/TCN, 2020. Accessed: April 2020.

[118] Google Apple. Privacy-preserving contact tracing. https://www.apple.com/covid19/contacttracing/, 2020. Accessed: April 2020.

[119] Bloomberg Technology. France says apple bluetooth policy is blocking virus tracker. https://www.bloomberg.com/news/articles/2020-04-20/france-says-apple-s-bluetooth-policy-is-blocking-virus-tracker, 2020. Accessed: April 2020.

[120] The Gaurdian. Nhs in standoff with apple and google over coronavirus tracing. https://www.theguardian.com/technology/2020/apr/16/nhs-in-standoff-with-apple-and-google-over-coronavirus-tracing, 2020. Accessed: April 2020.

[121] Regulation (eu) 2016/679.

[122] Private kit: Safe paths; privacy-by-design. https://safepaths.mit.edu/, 2020. Accessed: March 2020.

[123] The New York Times. In stores, secret surveillance tracks your every move. https://www.nytimes.com/interactive/2019/06/14/opinion/bluetooth-wireless-tracking-privacy.html, 2020. Accessed: March 2020.

[124] Forbes. The hidden trade-offs inside contact-tracing apps. https://www.forbes.com/sites/ramseyfaragher/2020/04/21/the-hidden-trade-offs-inside-contact-tracing-apps/#5a1e6700ea07, 2020. Accessed: March 2020.

[125] Bojan Nikolic. Enhancing bluetooth contact tracing for covid-19 containement with humidty/pressure sensors. http://www.bnikolic.co.uk/covid-19/bluetooth/2020/04/11/enhancing-bluetooth-proxity-covid19.html. Accessed: March 2020.

[126] Hyunghoon Cho, Daphne Ippolito, and Yun William Yu. Contact tracing mobile apps for covid-19: Privacy considerations and related trade-offs. *arXiv preprint arXiv:2003.11511*, 2020.

[127] Alex Berke, Michiel Bakker, Praneeth Vepakomma, Kent Larson, and A Pentland. Assessing disease exposure risk with location data: A proposal for cryptographic preservation of privacy. *arXiv preprint arXiv:2003.14412*, 2020.

[128] Robert Hinch, W Probert, A Nurtay, M Kendall, C Wymant, Matthew Hall, and C Fraser. Effective configurations of a digital contact tracing app: A report to nhsx. *en. In:(Apr. 2020). Available here. url: https://github. com/BDI-pathogens/covid-19_ instant_ tracing/blob/master/Report*, 2020.

[129] Ramesh Raskar, Isabel Schunemann, Rachel Barbar, Kristen Vilcans, Jim Gray, Praneeth Vepakomma, Suraj Kapa, Andrea Nuzzo, Rajiv Gupta, Alex Berke, et al. Apps gone rogue: Maintaining personal privacy in an epidemic. *arXiv preprint arXiv:2003.08567*, 2020.

[130] Sourabh Niyogi, James Petrie, Scott Leibrand, Jack Gallagher, Hamish, Manu Eder, Zsombor Szabo, and George Danezis. Cen proposals for privacy-preserving distributed contact tracing. `https://docs.google.com/document/d/1f65V3PI214-uYfZLUZtm55kdVwoazIMqGJrxcYNI4eg`, 2020. Accessed: March 2020.

[131] Sourabh Niyogi, Scott Leibrand, and Dana M. Lewis. Cen proposals for privacy-preserving distributed contact tracing. `https://github.com/Co-Epi/coepi-backend-go`, 2020. Accessed: March 2020.

[132] Günter Kampf, Daniel Todt, Stephanie Pfaender, and Eike Steinmann. Persistence of coronaviruses on inanimate surfaces and its inactivation with biocidal agents. *Journal of Hospital Infection*, 2020.

[133] Team Trace Together. Can i say no to uploading my tracetogether data when contacted by the ministry of health? `https://tracetogether.zendesk.com/hc/en-sg/articles/360044860414-Can-I-say-no-to-uploading-my-TraceTogether-data-when-contacted-by-the-Ministry-of-Health-`, 2020. Accessed: March 2020.

[134] Yves-Alexandre De Montjoye, Laura Radaelli, Vivek Kumar Singh, et al. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 347(6221):536–539, 2015.

[135] Yuval Noah Harari. The world after coronavirus. *Financial Times*, March 2020.

[136] Oliver Holmes. Israel to track mobile phones of suspected coronavirus cases. *The Guardian*, March 2020.

[137] Shui-yin Sharon Yam. Coronavirus and surveillance tech: how far will gov'ts go and will they stay when they get there? *Hong Kong Free Press*, March 2020.

[138] Benny Pinkas, Thomas Schneider, Gil Segev, and Michael Zohner. Phasing: Private set intersection using permutation-based hashing. In *24th {USENIX} Security Symposium ({USENIX} Security 15)*, pages 515–530, 2015.

[139] Pierre Baldi, Roberta Baronio, Emiliano De Cristofaro, Paolo Gasti, and Gene Tsudik. Countering gattaca: Efficient and secure testing of fully-sequenced human genomes (full version). *arXiv preprint arXiv:1110.2478*, 2011.

[140] Daniel Kales, Christian Rechberger, Thomas Schneider, Matthias Senker, and Christian Weinert. Mobile private contact discovery at scale. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pages 1447–1464, 2019.

[141] Ágnes Kiss, Jian Liu, Thomas Schneider, N Asokan, and Benny Pinkas. Private set intersection for unequal set sizes with mobile applications. *Proceedings on Privacy Enhancing Technologies*, 2017(4):177–197, 2017.

[142] Whitfield Diffie and Martin Hellman. New directions in cryptography. *IEEE transactions on Information Theory*, 22(6):644–654, 1976.

[143] Bernardo A Huberman, Matt Franklin, and Tad Hogg. Enhancing privacy and trust in electronic communities. In *Proceedings of the 1st ACM conference on Electronic commerce*, pages 78–86, 1999.

[144] Ran Canetti, Ari Trachtenberg, and Mayank Varia. Anonymous collocation discovery:taming the coronavirus while preserving privacy, 2020.

[145] Juanjuan Zhang, Maria Litvinova, Yuxia Liang, Yan Wang, Wei Wang, Shanlu Zhao, Qianhui Wu, Stefano Merler, Cecile Viboud, Alessandro Vespignani, et al. Age profile of susceptibility, mixing, and social distancing shape the dynamics of the novel coronavirus disease 2019 outbreak in china. *medRxiv*, 2020.

[146] Stephen M Kissler, Christine Tedijanto, Marc Lipsitch, and Yonatan Grad. Social distancing strategies for curbing the covid-19 epidemic. *medRxiv*, 2020.

[147] Michael L Jackson, Gregory R Hart, Denise J McCulloch, Amanda Adler, Elisabeth Brandstetter, Kairsten Fay, Peter Han, Kirsten Lacombe, Jover Lee, Thomas Sibley, et al. Effects of weather-related social distancing on city-scale transmission of respiratory viruses. *medRxiv*, 2020.

[148] Robert J Glass, Laura M Glass, Walter E Beyeler, and H Jason Min. Targeted social distancing designs for pandemic influenza. *Emerging infectious diseases*, 12(11):1671, 2006.

[149] Peter Caley, David J Philp, and Kevin McCracken. Quantifying social distancing arising from pandemic influenza. *Journal of the Royal Society Interface*, 5(23):631–639, 2008.

[150] Ulf Aslak and Laura Alessandretti. Infostop: Scalable stop-location detection in multi-user mobility data. *arXiv preprint arXiv:2003.14370*, 2020.

[151] M Bakker, A Berke, M Groh, AS Pentland, and E Moro. Effect of social distancing measures in the new york city metropolitan area. 2020.

[152] The New York Times. Location data says it all: Staying at home during coronavirus is a luxury. `https://www.nytimes.com/interactive/2020/04/03/us/coronavirus-stay-home-rich-poor.html`, 2020. Accessed: April 3 2020.

[153] Shengjie Lai, Nick W Ruktanonchai, Liangcai Zhou, Olivia Prosper, Wei Luo, Jessica R Floyd, Amy Wesolowski, Mauricio Santillana, Chi Zhang, Xiangjun Du, et al. Effect of non-pharmaceutical interventions to contain covid-19 in china. 2020.

[154] M Keith Chen, Yilin Zhuo, Malena de la Fuente, Ryne Rohla, and Elisa F Long. Causal estimation of stay-at-home orders on sars-cov-2 transmission. *arXiv preprint arXiv:2005.05469*, 2020.

[155] Parker Liautaud, Peter Huybers, and Mauricio Santillana. Fever and mobility data indicate social distancing has reduced incidence of communicable disease in the united states. *arXiv preprint arXiv:2004.09911*, 2020.

[156] Hanming Fang, Long Wang, and Yang Yang. Human mobility restrictions and the spread of the novel coronavirus (2019-ncov) in china. Technical report, National Bureau of Economic Research, 2020.

[157] Jayson S Jia, Xin Lu, Yun Yuan, Ge Xu, Jianmin Jia, and Nicholas A Christakis. Population flow drives spatio-temporal distribution of covid-19 in china. *Nature*, pages 1–5, 2020.

[158] Paolo Cintia, Daniele Fadda, Fosca Giannotti, Luca Pappalardo, Giulio Rossetti, Dino Pedreschi, Salvo Rinzivillo, Pietro Bonato, Francesco Fabbri, Francesco Penone, et al. The relationship between human mobility and viral transmissibility during the covid-19 epidemics in italy. *arXiv preprint arXiv:2006.03141*, 2020.

[159] Brendon Sen-Crowe, Mark McKenney, Dessy Boneva, and Adel Elkbuli. A state overview of covid19 spread, interventions and preparedness. *The American Journal of Emergency Medicine*, 2020.

[160] Jonas Dehning, Johannes Zierenberg, F Paul Spitzner, Michael Wibral, Joao Pinheiro Neto, Michael Wilczek, and Viola Priesemann. Inferring change points in the spread of covid-19 reveals the effectiveness of interventions. *Science*, 2020.

[161] Martin Andersen. Early evidence on social distancing in response to covid-19 in the united states. *Available at SSRN 3569368*, 2020.

[162] Rahi Abouk and Babak Heydari. The immediate effect of covid-19 policies on social distancing behavior in the united states. *Available at SSRN*, 2020.

[163] Hao Hu, Karima Nigmatulina, and Philip Eckhoff. The scaling of contact rates with population density for the infectious disease models. *Mathematical biosciences*, 244(2):125–134, 2013.

[164] Lara Goscé, David AW Barton, and Anders Johansson. Analytical modelling of the spread of disease in confined and crowded spaces. *Scientific reports*, 4:4856, 2014.

[165] Daxin Tian, Chao Liu, Zhengguo Sheng, Min Chen, and Yunpeng Wang. Analytical model of spread of epidemics in open finite regions. *IEEE Access*, 5:9673–9681, 2017.

[166] Javier Gutierrez and Juan Carlos García-Palomares. New spatial patterns of mobility within the metropolitan area of madrid: towards more complex and dispersed flow networks. *Journal of transport geography*, 15(1):18–30, 2007.

[167] Genevieve Giuliano and Dhiraj Narayan. Another look at travel patterns and urban form: the us and great britain. *Urban studies*, 40(11):2295–2312, 2003.

[168] Becky PY Loo and Alice SY Chow. Changing urban form in hong kong: what are the challenges on sustainable transportation? *International Journal of Sustainable Transportation*, 2(3):177–193, 2008.

[169] The New York Times. Coronavirus (covid-19) data in the united states. `https://github.com/nytimes/covid-19-data`. Accessed: June 2020.

[170] U.S. Centers for Disease Control. *COVID-19 Travel Recommendations by Country*, 2020 (accessed June, 2020).

[171] Solomon Hsiang, Daniel Allen, Sébastien Annan-Phan, Kendon Bell, Ian Bolliger, Trinetta Chong, Hannah Druckenmiller, Luna Yue Huang, Andrew Hultgren, Emma Krasovich, et al. The effect of large-scale anti-contagion policies on the covid-19 pandemic. *Nature*, pages 1–9, 2020.

[172] U.S. Centers for Disease Control. *Contact Tracing Resources*, 2020 (accessed June, 2020).

[173] Justin Chan, Shyam Gollakota, Eric Horvitz, Joseph Jaeger, Sham Kakade, Tadayoshi Kohno, John Langford, Jonathan Larson, Sudheesh Singanamalla, Jacob Sunshine, et al. Pact: Privacy sensitive protocols and mechanisms for mobile contact tracing. *arXiv preprint arXiv:2004.03544*, 2020.

[174] Hao Chen, Kim Laine, and Peter Rindal. Fast private set intersection from homomorphic encryption. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1243–1255, 2017.