Big Location Data:
Balancing Profits, Promise, and Perils

Thesis Proposal

Chris Riederer
Department of Computer Science
Columbia University
mani@cs.columbia.edu

April 18, 2017

Abstract

The “Big Data" era has begun, bringing with it a host of possibilities and concerns. The ability to store and process records of minute behavioral details about billions of people will hopefully lead to more efficient and effective businesses, governments, and organizations. At the same time, these new collections of data present societal risks, enabling mass surveillance, a potential loss of privacy, and the capability to computationally discriminate at massive scale.

A rich subset of this data is human mobility data: records at an individual scale detailing where and when someone moves. This data is now captured like never before due to the rise of smartphones and other cheap and ubiquitous electronic devices. This location data can be a boon to both profit centers and scientific understanding, but comes with many risks attached. The places we visit can reveal much about ourselves, whether proclivities towards particular type of food and hobbies, or more private characteristics of race, religion, sexuality, and political affiliation. As organizations begin to harness this data, it is clearly important to make sure that such information is used in a way that reduces potential harms to individuals or vulnerable groups.

In this document, I describe recent and in-progress work that attempts to balance the scientific and engineering promises of location data with the potential risks, looking at three classes of problems. I begin with my work on anonymity, examining when location data retains an identifiable “fingerprint" of a user that can be linked to data sets generated by completely separate behaviors, making anonymization difficult or impossible. I continue with work focusing on privacy and economics, aiming to reconcile the economic incentives for firms to collect location data with usable user choice and a better understanding of users’ beliefs and desires in relation to data capture. Next, I examine the relationship between location data and algorithmic bias, showing that location data can be used to infer sensitive traits and developing a tool to inform users about what their data may be revealing.

I conclude with my proposal to conduct an analysis that will show the conditions in which a firm may engage in location-based advertising without unfairly distributing their offers across different demographics. This work will combine a dataset of millions of locations collected from social media, computer vision techniques, and state-of-the-art results in algorithmic de-biasing to show, for the first time, the trade off between revenue and fairness in a location-based advertising setting.

Contents

Chapter 1

Introduction

The era of Big Data has the potential to transform all aspects of society, from business operations to individual leisure time, from medicine to elections. As the cost of computation and data storage has fallen, and as more human behavior has moved to easily-recorded digital mediums, both public and private institutions have captured and stored more and more information. “Data is the new oil", analysts have proclaimed, as businesses seek to become more efficient or create innovative digital products. “Data will power the cities of tomorrow" governments have proclaimed, releasing their troves of data to the world.

The excitement of the potential gains has been tempered by many potential concerns. With the desire for public agencies that could more accurately prioritize safety concerns has come massive surveillance, both by companies and governments, raising concerns of **anonymity**. With the belief that Big Data would lower costs and benefit consumers has come evidence of large-scale price discrimination [33, 21], with its impacts possibly most felt by already economically vulnerable populations, raising concerns of **economics**. And with the hopes of making government more fair and efficient have come allegations that seemingly impartial algorithms actually encode unfair racial biases [2], raising concerns of **algorithmic bias**. We must carefully consider how to obtain the benefits of the Big Data Age without paying too high a price.

One particularly interesting and sensitive subset of Big Data is information relating to the real world movements of individuals: “location data". As this data is tied to actual physical movements, it is of enormous interests to businesses and governments. When people move, they are expending energy and money to do so, as opposed to the many relatively low-cost digital behaviors captured online. Thus, location data provides a valuable input for many problems, as evidenced by academic proof-of-concepts and by the many businesses and services surrounding it’s capture, such as Google Location Services, Foursquare[1], Placed[2], and many more.

Along with its valuable signal again comes concerns of safety and privacy. In the United States v. Jones decision, in which the Supreme Court wrestled with the legality of law-enforcement placing a GPS tracker on a car without a warrant, Justice Sotomayor wrote “disclosed in [GPS] data ... [are] trips the indisputably private nature of which takes little imagination to conjure: trips to the psychiatrist, the plastic surgeon, the abortion clinic, the AIDS treatment center, the strip club, the criminal defense attorney, the by-the-hour motel, the union meeting, the mosque, synagogue or church, the gay bar and on and on" [31]. Clearly, if location data is to be used by analysts, it must

[1] `foursquare.com`
[2] `placed.com/`

be done so in a way that is extremely careful and respectful of user preferences.

In this thesis proposal, I examine work that attempts to reconcile the benefits of data-mining location data with three key areas of concern: anonymity, economics, and algorithmic bias. I will begin with a background section which introduces the core concepts found in this proposal, continue with several chapters relating to the afore-mentioned areas of concern, and conclude with a final discussion of next steps in making big location data safe and fair to use.

1.1 Anonymity

Chapter 2 examines the possibility of anonymizing location data. Prior work has shown that users are highly unique in their location patterns, leaving them vulnerable to de-anonymization. Here we take this a step further, showing not only that this vulnerability exists, but that users indeed can be linked to other datasets. Additionally, we provide a tool to users that aggregates and displays their location data along with the potential inferences made from it.

1.2 Economics

I focus on location data, privacy, and economics in Chapter 3. We begin with work that seeks to understand user attitudes to their privacy and the economic value of their information. Specifically, it examines an alternative to the current practice of firms offering free services in exchange for full control over user data. The alternative model is one in which users control their data and make decisions about when to sell access to their info, and to whom.

1.3 Algorithmic Bias

Finally, I analyze the interaction between location data and algorithmic bias in Chapter 4. We gather a dataset of locations attached to demographic information from a popular image-sharing mobile application. This data allows us to study the differences in human mobility across different groups, and moreover, to show that demographics can be inferred using only location data. This raises questions about the sensitivity of location data, and about the potential for bias in systems that make decisions based on location data. We examine other methodologies for inferring demographics from social network data and discuss de-biasing of algorithms. Furthermore, I demonstrate a personal auditing tool for users to understand their location data and the inferences potentially made about them.

1.4 De-biasing Location Based Advertising

Chapter 5 links my previous works together in my proposal for future work. I hope to discover the conditions under which advertisers may safely use location data to decide when to show ads without unfairly benefiting one demographic group at the cost of another. Advertisers (and other users of algorithms) often wish to target users based on certain traits (e.g. interest in a product). These users may be clustered geographically, for example, only users in New York will be interested in

purchasing a subway pass for that city. While wishing to target, advertisers may also wish to *not* exclude certain demographics due to issues of fairness (and potential bad publicity). Our New York City subway pass advertiser might not wish to offer more deals to the wealthy at the expense of the poor, for example. By using a large dataset of location data tagged with user generated text and demographic information, I will be able to determine the cost of altering a location-based advertising algorithm to insure that ads are shown to a representative population. A key metric will be demographic mobility, looking at how skewed from the overall distribution the make up of visits by certain demographics are in a location. In locations where the difference in demographic mobility is small, advertisers can tweak their algorithms to show ads equally across demographics with minimal cost, fair to both individuals and overall groups. In other cases where the difference is large, the costs will likely be higher. I offer more detailed description of this planned work in Chapter 5.

1.5 Background

What is location data? Most generally, location data is information relating people to places. Typically, this relation is the fact that a person was at a place. Adding time into the figure, the relation could be that a person was at a place at a particular time. However, location data could also include relations about the importance of a place in someones life, such as them living in a location, working at a location, or having spent a quantity of time in a location. Though location data does not need to be associated with user IDs, in this work we will consider that there is always attached some sort of user ID that uniquely identifies the user in the dataset, possibly de-personalized.

Location data can be described in two main ways: **geographically** or **semantically**. *Geographic* data can be described by a latitude-longitude data on the globe. *Semantic* location data refers to an identifier used within that dataset. This could have some information available to a common user, e.g. “New York City", or it could simply be an identifier, e.g. 7. Note that often these two may be combined or used together. A location such as “CEPSR Office 618, Columbia University" (the author’s office) indicates a very small, non-ambiguous location that can easily be mapped to geographic coordinates. Semantic location data can sometimes present a privacy problem, as an association with a place could indicate sensitive attributes, such as someone’s religion, political affiliation, health, or sexuality. In this work, I will typically assume location data is also tagged with temporal data, and I will use the terms location data and spatiotemporal data interchangeably.

To put this more formally, we can define a single data point p of location data to be:

$$p = \langle u, l \rangle$$

or, including time:

$$p = \langle u, l, t \rangle$$

where u uniquely identifies a user, l uniquely identifies a location, and t specifies a time. Note that l could be a latitude-longitude pair in the geographic case or an ID in the semantic case.

How is location data collected? Location data can be captured passively or actively. **Actively captured** location data is only recorded when the user takes some action. Note that this action does not need to inherently be “about" location data, for example, a user making a call from a cell phone

or swiping a credit card is typically not consciously thinking about their location data. A record of their location is created as a by-product of their use of that technology. **Passively captured** is meant in a stronger way– the user's location is captured without the user making any kind of action. This can occur through tracking apps. An example is MapMyRun[3], an app where users record their routes while running, in order to track distance and progress in meeting exercise goals. Although the user took an action to start recording their location, the location is recorded in the background with no user action from then on, and hence we call it "passive". Another example is Google's location history. Google records location data in the background of a users Android phone every few minutes. A map of everywhere a user (with an Android phone with location history turned on) is available at [4].

[3] http://www.mapmyrun.com/

[4] https://www.google.co.in/maps/timeline

Chapter 2

Location Data and Anonymity

Location data is extremely sensitive but also useful for many applications. In his survey of computational location privacy [19], Krumm lists four main methods of protecting location privacy while still using location data:

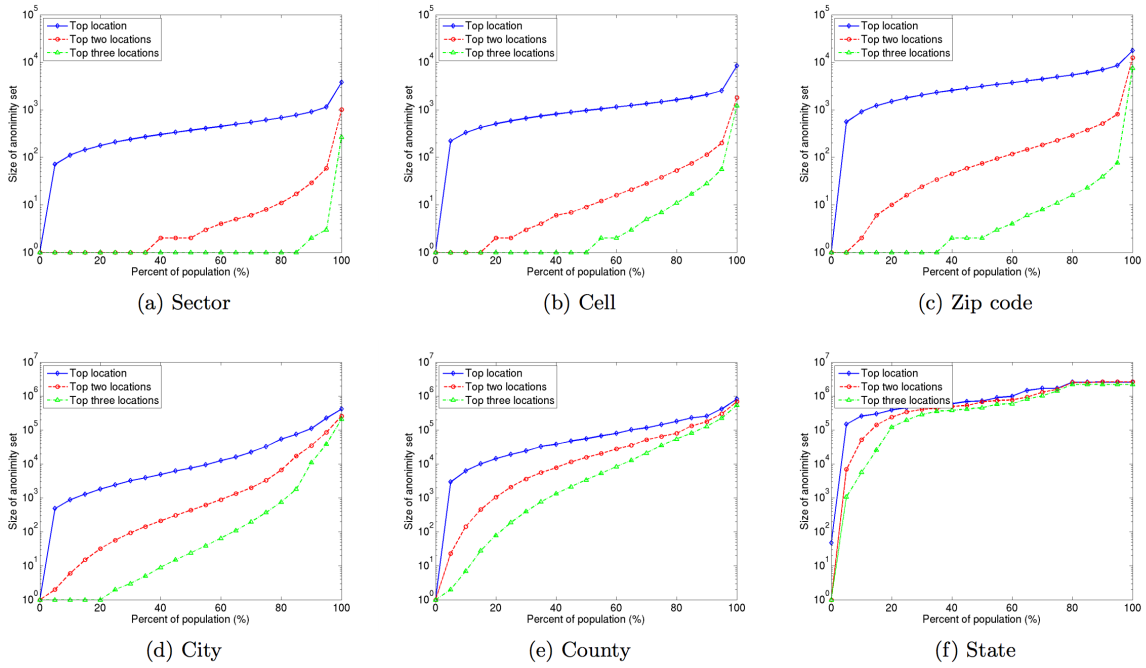
1. Anonymity
2. Obfuscation
3. Regulatory strategies
4. Privacy policies

Anonymizing a user ID associated with location data may sound good in theory, but due to the fact that a typical users movements are very unique, many attacks can re-identify these users in practice. This raises the question: when can anonymized location data be used or released without fear of de-anonymization? In this section, we explore the difficulties in anonymizing location data.

2.1 Related Work

Location data for individuals is highly unique and thus difficult to anonymize. The first large-scale study of the k-anonymity of location data was appropriately titled “Anonymization of Location Data Does Not Work" [35]. The paper used data from cell phone call detail records (or CDR, see Chapter 1.5) for 25 million United States users over a 3 month period. The authors represents each user as simply their top n most visited locations, varying n from 1 to 3. Additionally, the authors varied the granularity of the locations, with the smallest as cell sector and the largest as state. Remarkably, using 3 locations at a cell level made half of all users completely unique, and 3 locations a sector level made 85% of all users unique. A figure detailing this result and results for other granularities and values of n is depicted in Figure 2.1. The authors went on to analyze the impact of geography (comparing different states and cities), mobility (distances between top locations), and social networks on anonymity.

A different study used randomly selected points from a user’s dataset and included time of location visit [8], as opposed to a users top n locations (mostly omitting precise time) of Zang and Bolot. Using a call detail record dataset of 1.5 million users from a small European country, this



(a) Sector (b) Cell (c) Zip code

(d) City (e) County (f) State

Figure 2.1: Figure from [35] depicting the size of anonymity sets for top n most visited location of users. Locations are varied in granularity, from cell sectors to US states.

work showed that 95% of users are uniquely identified by 4 spatiotemporal points. A follow up study [9] showed that in a data set of credit card transactions, user profiles of spatiotemporal points had a similar level of uniqueness, and even more when transaction amounts were included as well.

2.2 Linking Users Across Domains with Location Data

Although prior work showed location to be highly *unique* and thus possibly *vulnerable* to de-anonymization, no data was actually de-anonymized in practice. Indeed, just because a data source is highly unique does not mean it can be de-anonymized. To put it more concretely, imagine that each individual had a die with 1000 sides, and each side represented a location. For example, much of cryptography relies on creating unique but unpredictable sequences of numbers. If, quite hypothetically, humans decided where to go next by rolling this die, their movements would look very unique. However, since the movements are random and unpredictable, my movements from different time periods will be indistinguishable from those of a different individual.

Another possible break in the argument that uniqueness implies vulnerability is the important factor of sampling. The datasets dealt with here (phone records, social media posts) are all *actively* collected: each data point exists if and only if the user has taken an action. Intuitively, the location data from different sampling data sources should look very different. An individual may be more likely to make phone calls in quiet places, like the home or office, and take geotagged location photos in popular tourist destinations or restaurants.

The vector of attack implied in the above-mentioned works are having access to an origi-

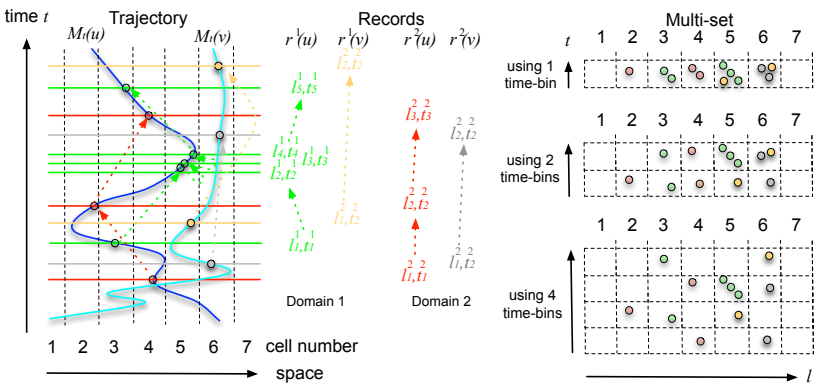


Figure 2.2: Two space-time trajectories with associated footprints in two domains.

nal, anonymized dataset, and then de-anonymizing it with a subset of auxiliary information with columns identical to the original set. In “Linking Users Across Domains with Location Data", published at WWW in 2016 [26], we tackled the problem of linking users across two entirely different datasets using only their location data.

We formalized the problem in the following manner. We defined U and V to be sets of n user accounts in two separate domains. Each account is a itself a set of spatiotemporal points p, where

$$p = \langle u, l, t \rangle$$

with u being a user ID unique to either U or V, l is a location, and t is a time. We denoted σ_I to be a true (“identity”) mapping that correctly links the two accounts of the each user across U and V. The goal then, of this work, is to recover σ_I.

We made a series of simple assumptions about human mobility. We broke time into discrete “bins" of a certain length, and then declared the number of checkins a user has at each location in time bin to be Poisson distributed according to a rate parameter λ unique to that time and place. This is a simple but reasonable assumption, and Poisson distributions are often used to model rare events (like checkins).

This model generates the *real world* mobility of a user. We assume that this real world mobility is sampled independently and randomly for the two different data sets with probability p_U and p_V.

Figure 2.2 provides a visual illustration. On the left side of the image are two real world trajectories, denoted with a blue and turquoise line. The x axis shows space and the y axis shows time. The colored circles (red, green, gray and yellow) show times and places where the real world trajectories are sampled, with (for example) a geolocated photograph, phone call, or checkin. The challenge is that we only see the green, yellow, red, and gray trajectories in the middle of the image, and we must figure out the true association across datasets. In this example, red should go with green and gray with yellow. On the right side of the image, the concept of time bins are illustrated. We discretize time with varying sized time bins. The top uses one large time bin, essentially ignoring time, whereas the bottom breaks time into four sections, essentially saying two locations are only the same if the checkins occur near one another in time.

We evaluated this algorithm on multiple real-world datasets. Gathering the data in itself was a significant challenge, as each dataset needed to contain individuals with identities linked across two different data sources. Collecting information from one data source is enough of a challenge

Dataset	Domain	Number Users	Number Checkins	Median Checkins	Number Locations	Date Range
FSQ-TWT	Foursquare	862	13,177	8	11,265	2006-10 – 2012-11
	Twitter	862	174,618	60.5	75,005	2008-10 – 2012-11
IG-TWT	Instagram	1717	337,934	93	177,430	2010-10 – 2013-09
	Twitter	1717	447,366	89	182,409	2010-09 – 2015-04
Call-Bank	Phone Calls	452	∼200k	∼550	∼3500	2013-04 – 2013-07
	Card Transactions	452	∼40k	∼60	∼3500	2013-04 – 2013-07

Table 2.1: Overview of datasets used in study. For FSQ-TWT and IG-TWT, number of locations refers to locations at a 4 decimal GPS granularity (position within roughly 10m).

by itself, given unexpected and changing data formats, connectivity problems, rate limits, and more. Getting ground truth data across two datasets is thus more difficult, as two APIs need to be dealt with and user identities must be verified across the two.

We gathered three datasets:

- **Foursquare-Twitter** (FSQ-TWT): checkin data from the location-based social networking and review site Foursquare [1] and geotagged updates from the microblogging site Twitter [2]. This data was obtained in a prior work by other authors who allowed us to use their data [36]. We expect the behavior to be somewhat different across the two networks; Foursquare is primarily used to review restaurants, and Twitter is generally used.

- **Instagram-Twitter** (IG-TWT): Geolocated photographs from the image sharing site Instagram [3] and geotagged updates from the microblogging site Twitter. We first crawled Instagram, and then found users who had posted their Twitter usernames in their profiles. For each of these users, we used Twitter's API to crawl their public tweets. We expected this dataset to be the easiest to link, as there were high numbers of checkins on both sites for most users.

- **Cell phone-Credit Card** (Call-Bank): Phone calls associated with geolocated cell towers (CDR) and credit or debit card transaction data associated with geocoded businesses, all from one G20 country. Locations were declared the same if the lat-lon of business was within a cell created via a Voronoi tesselation. This data was very sparse and the behaviors generating data seems to be very different in the two sets, making us hypothesize that we would have our worst results on it.

Statistics about these datasets is summarized in Table 2.1.

We now turn our attention to experimental performances of our algorithm. In Figure 2.3, we show the precision recall plots for our algorithm (for different eccentricity values) and for the other three reconciliation techniques: HIST, NFLX and WYCI, which are, respectively, the histogram matching technique of [32], the "Netflix De-Anonymization Attack" [22], and a frequency-based

[1] `https://foursquare.com/`

[2] `twitter.com`

[3] `instagram.com`

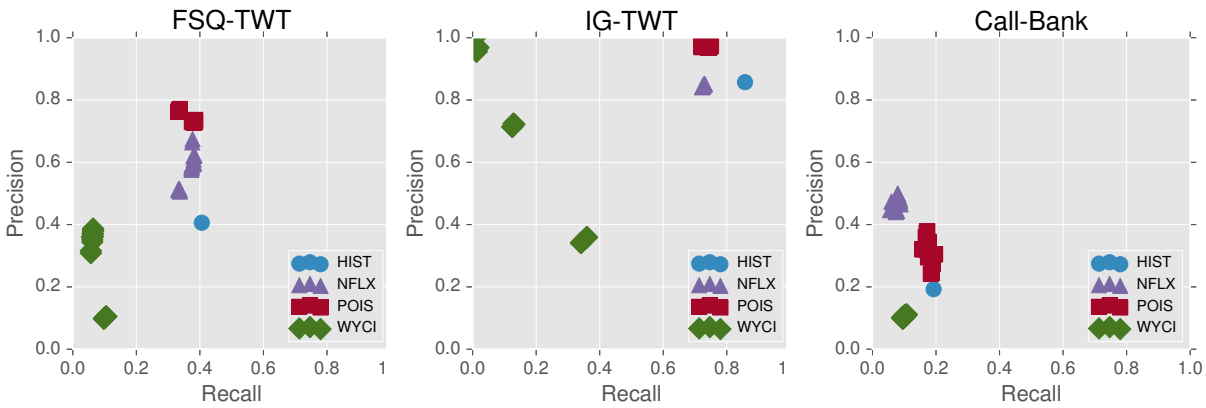


Figure 2.3: Precision and Recall plots for each dataset.

maximum likelihood estimator [29]. For our algorithm, we used estimated parameters and for the other techniques, we used optimal parameters (found via exhaustive search).

There are several interesting observations that we can make on Figure 2.3. First, on the public dataset FQ-TWT our algorithm outperforms all prior methods (especially in precision). Nevertheless it is interesting to note that the precision of all methods is not ideal, probably due to sparsity of the data.

A second interesting observation is that our algorithm achieves very high precision when the dataset is more rich. In fact when we then turn our attention to our second dataset, the live service (IG-TWT) that we crawled, we obtain almost perfect precision. Note that not all the other techniques, for example NFLX, are able to leverage the denser data, as much.

Finally we test our method on a much more heterogeneous dataset (Call-Bank) that is also more realistic and sensitive. In this setting our algorithm outperforms previous techniques, with none of the previous algorithms able to achieve good precision and recall at the same time.

Additional results fully described in the paper found that our algorithms rapidly improved with more data. Additionally, varying the size of timebins or the eccentricity parameter or number of terms did not have a large impact on results, meaning our algorithm's performance should remain stable to different sets of parameters.

Chapter 3

Location Data, Privacy, and Economics

The online attention economy is based primarily on advertising. Some of the most powerful and far-reaching online services, such as Google and Facebook, operate mostly with free services. The income of a firm roughly translates to (number of impressions) x (dollars per impression). Firms are not necessarily paid on an impression basis, as other models like paying per click or per sale or other action are quite common, but it is reasonable to think of clicks or sales as some function or subset of impressions. Fundamentally, much of the dollars in profits given to online firms are based on how many people come to their site and how well they can target advertisements to those individuals. This gives firms an incentive to gather as much information about their users as possible so that they can better target ads to them.

This framework presents a challenge to privacy. User information is collected and gathered in one centralized place. There are multiple risks involved here: the firms themselves may use the information in ways the users disagree with, the firms may sell or be coerced to give their information to other firms or governments, or the firms may fall victim to cybersecurity attacks, leaking information to other sources.

As ways to counter this, schemes have been proposed to encrypt user behavior and information, denying all access to a firm. However, this would deny firms the ability to make money, meaning no services would be provided for users and possibly a lower global utility be reached. Thus, schemes that ignore this economy however are unlikely to be adopted. Companies need to make money to function. Currently, users seem happy to provide their data in exchange for free services. A concern is that users do not have a good idea of their data and do not know how it is being used and to whom it is accessible.

Therefore it is important to gain an understanding of how users value their information, what they believe firms are doing with their data, and what users are comfortable with in terms of data use. Building upon my previous work analyzing how users value their information [7] and proposing alternative data discolure methods [25], in this section I propose a new method for representing user locations in a privacy sensitive manner that can still be effective for advertisers.

3.1 User Choice in Location Disclosure

Many privacy concerns around location information are rooted in the mobile application ecosystem. Most mobile services and applications are free and operate by collecting personal information

(browsing activity, location, etc.) and monetizing this information through targeted ads [20]. In fact, one may argue that users today exchange their data for services. An ideal privacy solution therefore should provide adequate privacy protection to the user while simultaneously enabling service providers to collect and monetize data. Our objective is to lay the groundwork for a comprehensive and deployable solution to location privacy.

In this section, I describe published work [28] that aims to reconcile the users' control over their location information with its commercial value. This approach raises three challenges: (1) The solution should be *incrementally deployable*. It must easily integrate with current devices and practices while giving all parties an incentive to participate. (2) The solution should be *robust* against threats from its participants. Advertisers should not be able to access data without compensating users or access more than the users specify. Users should not be able to benefit from seeking unfair compensation. (3) The solution should be *easy to use*. The system should be easily understood by both users and advertisers.

Our solution is based on selective disclosure; users decide what location information they want to disclose. At the heart of our solution is a *keyword-based* method where keywords are associated with locations, and the decision to release locations is based on keywords. We observe that keywords are naturally associated with the elements that define this problem, but also offer a strong abstraction to handle location data. In order to drive the adoption of the solution, we propose providing economic compensation to the users for the location information they disclose. Application and web service providers bid to gain *access* to users at these specific locations in real-time.

3.1.1 Overview

Our requirements calls for a solution to share information about location monetized by ad-networks and 3rd party aggregators through *selective disclosure*. For the user to retain control, our privacy solution should address *how* the information is released, under *which conditions* the information is released and to *whom*, as seen in previous ones, e.g. Koi [12],

To specify *how* and *under which conditions* location information is released, we choose to use keywords. While the information that is released is a latitude longitude pair (lat-long), the decision to disclose is based on associated keywords. Users who are comfortable disclosing location under certain circumstances [17] opt-in to reveal lat-long associated with keywords of their choices. An example would be a street that has many restaurants serving different cuisines, it would have keywords like "restaurant, Thai, French, Indian" associated each with the lat-long of each particular venue. The use of keywords brings important advantages: (i) Keywords let us deal with the problem of location privacy at a higher abstraction than coordinates or even location descriptors as in Koi [12]. (ii) Keywords are user friendly: instead of having to decide the sensitivity of every location, users decide on a much smaller set of keywords that they are comfortable releasing or not. (iii) Today's ad-networks function primarily around keywords, thereby a solution around keywords can make it easier for ad-networks to adopt and use. (iv) As there can be a finite set of keywords associated with any location, and the association of a keyword with a location typically remains for long periods of times, modifying keywords associated with a location is easy, making the solution scalable.

Our solution compensates users *economically* for information they release to aggregators and ad-networks. Economic incentives can nudge more users towards adoption, as concerns about privacy alone are rarely sufficient. Concrete incentives also sometimes reduce users' cognitive

biases when it comes to perceiving their privacy [5]. Specifying to *whom* the information is released is implicitly done by a market. In principle, any parties that can pay for it is legitimate. In practice, this agreement should be facilitated by a trusted third party who vet the parties and send information about the user *only* for locations she agreed on, upon payment.

The architecture consists of the following components: (i) a keyword server which maps physical locations to keywords (ii) a location blacklist module which contains a list of sensitive keywords, communicates with the keyword server, and reveals non-sensitive locations (iii) a blocking module in the network that blocks access to various parties, (iv) a market that puts up for sale information about locations visited by the user that are not in the blacklist, and (v) a module that grants *access* to the user for parties that pay, after purchasing access on the market. With the exception of (ii), which can be a simple smartphone app, all modules are stored in the network; *no* changes are required on the device.

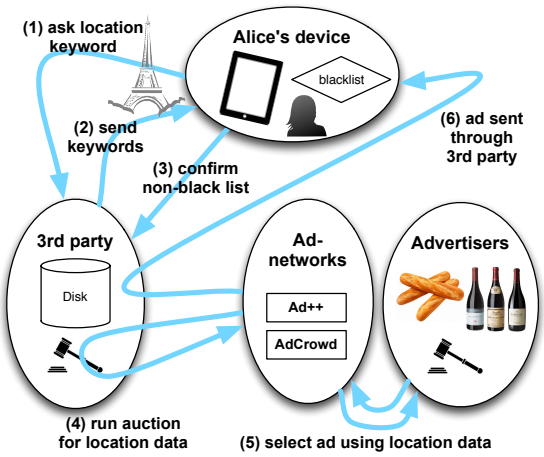


Figure 3.1: Solution overview

A high-level diagram is shown in Fig. 3.1. We describe the process with a simple example. Alice is willing to share certain locations and would like to hide her presence at other locations, a typical occurrence [17]. Alice wants to buy bread, shop for wine, and go to the Libertarian party headquarters. She would like to conceal her political leanings. Alice would therefore put 'Libertarian, Politics' as keywords in her *blacklist module*. We assume the third party is trusted and leave lowering this requirement to future work.

3.1.2 Deployment and User Study

An implementation consists of the five components: a keyword server, a location blacklist module, a network blocking module, an information market, and an access module.

Our **keyword server** used Yelp's API. Each time a device uploaded a lat-long to the server, we queried Yelp to find the categories of each location within 50 meters, using these categories as the location's keywords. The **location blacklist** module was written as an Android application. The app was designed to give users a way to edit a blacklist and monitor which locations (and corresponding keywords) were being recorded. We used Yelp's 885 categories as our keywords during the study, meaning users had a large number of potential keywords to blacklist. To make adding keywords to the blacklist manageable, all possible keywords were placed in a nested menu by category. We placed categories previously defined to be sensitive [3] near the top of this list, and alphabetized all potentially less sensitive categories. The blacklist was stored locally on the phone. *At no point did the authors have access to a study participant's blacklist.* Each half hour, the app would passively check the keywords in the current location and upload the location and keywords to the server only if no keywords were on the blacklist. For the purposes of our small scale user study, we did not create a **blocking module** or connect the system to any ad exchanges.

Instead of implementing a **market** or **access module**, we simulated the incentives and costs a user might experience while using our system. In our deployment, all participants received $10 for participating and were entered into a lottery. Each user was instructed that releasing more 'valuable' information would give them a higher chance of the lottery. We did not disclose the exact method of valuing information, mimicking the opaque way in which information would be priced in a real implementation of the system. The intention was that this would incentivize users to release more information. To protect users' safety, users could contact us at any point if they were concerned about an unintentional location release. Additionally, any time a data point was recorded, we delayed making it public by 24 hours. Users could see their data points in real-time via a password-secured link.

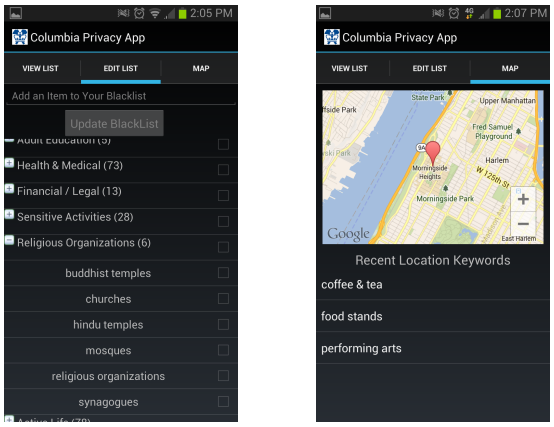


Figure 3.2: User Interface: (left) managing keywords black list, (right) visualizing locations released.

We deployed our implementation with six users for two weeks. Users were geographically diverse, located in multiple cities throughout the United States. Study participants were recruited through advertising on social networks and were primarily adults in their mid-twenties.

After the study, we asked users to complete a survey. Our study was too small to make general conclusions, but we present results here to inform future work. Users easily understood both the keyword system and the interface. Users were divided on how well they felt the system secured their privacy, with some users concerned that our mapping of keywords to locations was not precise enough. Our users expressed a range of privacy sensitivities. Some did not use the blacklist and others used the blacklist to hide sites they associated with social stigma or that they thought would send negative signals to employers, insurers, or the police.

Chapter 4

Location Data, Demographics, and Bias

4.1 Demographic Mobility

Human mobility is intimately intertwined with highly personal behaviors and characteristics. Previous studies of mobility centered on the risk of either re-identification in sensitive anonymized location datasets or on protecting visits to private locations [8, 13]. However, the re-identification risk based on individual locations is not the only threat. In this section, the full version of which appears in [27], we explore the discriminative power of location data. Solely based on mobility patterns, which we extracted from photosharing network profiles, we infer users' ethnicities and gender both on a demographic and an individual level. This exploration stands in contrast to limitations of previous studies as our paper brings together the following contributions:

- We show how photosharing network data can be leveraged to extract mobility patterns using a new method for creating location datasets from publicly available resources. Our method combines the use of online social networks and crowdsourcing platforms. It has the advantage that it generally enables *anyone* to study human mobility and does not mandate access to Call Detail Records (CDRs) or other proprietary datasets.

- To assess the quality of the created datasets we show that mobility patterns extracted from photosharing networks are comparable in terms of their essential characteristics to those previously observed and reported for CDRs. For the first time, we extend the analysis of mobility patterns to *ethnic groups*. We show how comparisons lead to statistically significant differences that are meaningful for assessing residential and peripatetic segregation.

- Finally, we demonstrate the discriminative power of location data on an *individual* level. Our analysis confirms for the first time that location data alone suffices to predict an individual's ethnicity, even with relatively simple frequency-based algorithms. Moreover, this inference is robust: a small amount of location records at a coarse grain allows for an inference competitive with more sophisticated methods despite of data sparsity and noise.

User profiles on photosharing networks often contain a significant amount of photos tagged with latitude-longitude GPS locations. Over time the accumulated location data can build up to comprehensive mobility profiles. Based on this insight and given that many user profiles on photosharing networks are publicly accessible we now introduce a methodology and its application to

construct mobility datasets from readily available data. An overview of our methodology is shown in Figure 4.1. This methodology is expanded and studied in more detail in Chapter 4.2.

Data Collection. We collected publicly available photo metadata from Instagram covering data for the years from 2011 through 2013. This data collection and use was exempt from user informed consent under our institution's IRB rules since (1) we only collected publicly available online metadata, (2) after we used the metadata and the users were labeled, any identifying information, such as usernames, were removed, and (3) we only kept track of users' identities separately and for one single purpose (ensuring that the data we collected still belongs to a public Instagram profile). We started our crawl from a root user (the founder of Instagram, on whose feed a large and diverse group of users comment) and followed further users subsequently through comments and likes. We skipped users with no geotagged photo in their first 45 photos. Our crawl retrieved a total of 35,307,441 photo location points belonging to 118,374 unique users.

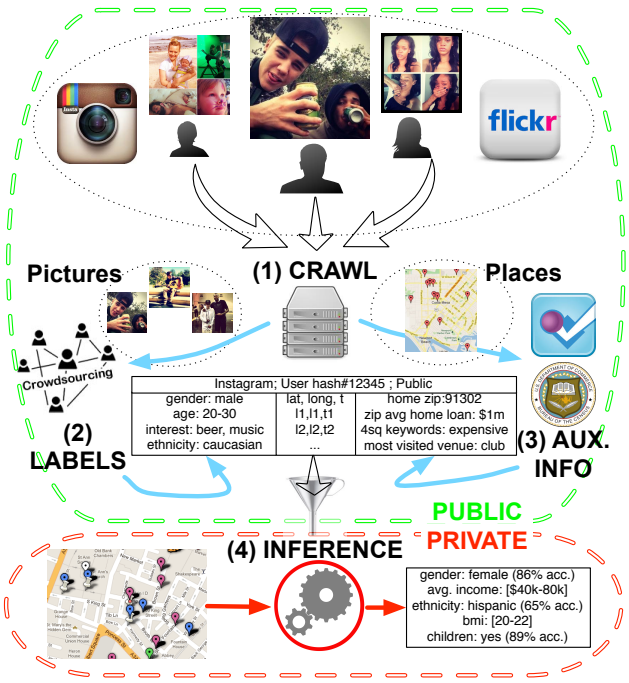


Figure 4.1: Methodology overview.

User Labeling. We labeled users for gender and ethnicity, first doing a manual comparison of trusted labelers before scaling with crowdsourcing. We selected a subset of profiles to label, being unable to label the entire dataset due to cost. In order to compare our data set to previous studies [14, 15, 16], we selected users from the Los Angeles (LA) and New York City (NY) metropolitan areas, based on a majority of a user's locations being in this area to filter out tourists. We ran a preliminary experiment by selecting 200 profiles at random (excluding celebrities and business accounts) and labeling each independently by two undergraduate students. We observed a strong agreement on gender (98%). For ethnicity labeling we leveraged United States Census categories, asking the student annotators to categorize each user either as Hispanic or Latino (Hispanic), White alone (Caucasian), Black or African American alone (African American), or Other (combining all remaining Census categories, including Asian). Just as in the Census, our Hispanic category includes Hispanics and Latinos of any race, while the remaining categories do not include any Hispanics or Latinos. We found that our profiles are diverse: 45% Caucasian, 21% Hispanic, 15% African American, and 19% Other. The students' labels matched 87% of the time and when evaluated as a binary classification task (Caucasian vs. all other categories) the agreement reached 94%. We note that the two labeling students were of different gender and ethnicity themselves.

We scaled our annotation using Amazon Mechanical Turk workers, with each profile labeled by two MTurk workers. In cases of disagreement between the MTurk annotators we asked one of our undergraduate annotators for an additional label to break the tie or assign a label from a different third category. We decided to use a tiered annotation mechanism with the undergraduate annotator

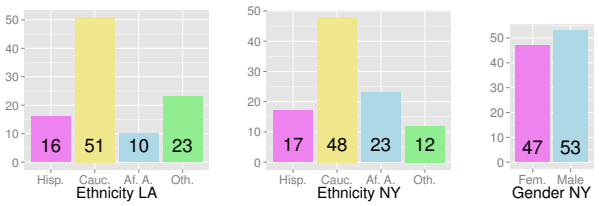
making the final decision in case of disagreements. We ran this on two different days, with workers paid $0.10 per profile on the first day and $0.05 on the second. This resulted in 1,015 labeled profiles. The undergraduate annotator was compensated the regular stipend at our institution. In order to measure the quality of agreement among the annotators we made use of Krippendorff's α [18], as shown in Figure 4.2. We collected auxiliary information mapping locations to the race and gender of residents using the United States Census [30] and location to categories using Foursquare's reverse-geocoding API.

4.1.1 Inferring Demographics from Locations

Task	*Best Algorithm*	*Parameters*	*Important Features*	*Baseline*	*Accuracy*	*AUC*	*F1*
Ethnicity NY	Log. Regression	L1, $C = 0.01$	Avg. ZIP ethnicities	0.52	**0.72**	**0.76**	**0.74**
Ethnicity LA	Log. Regression	L1, $C = 1$	Avg. ZIP ethnicities	0.50	**0.63**	**0.66**	**0.64**
Gender NY	Log. Regression	L2, $C = 0.1$	Men's Store	0.53	**0.58**	**0.59**	**0.55**

Table 4.1: Results for the binary classifications of ethnicity and gender in NY and LA. The algorithms ran on all available features, such as counts of visits to different neighborhoods, the ethnicity of the most visited block, and the categories of nearby Foursquare venues. The baseline was obtained by predicting the class of a user based on the label distribution.

We now show how location data by itself allows to infer ethnicity and gender of individual Internet users. Our purpose is to explore generally what might be inferred about users from their location data only. This means we utilized only well-understood, and commonly applied techniques, and limited our feature sets to features derived from location data (omitting, for example, social network features like number of followers). We considered two questions: (1) Can minorities be distinguished from Caucasians? (2) Can women be distinguished from men?



	Ethnicity LA	*Ethnicity NY*	*Gender NY*
Users (n)	427	588	241
K.'s α Multi.	**0.74**	**0.68**	-
K.'s α Binary	**0.78**	**0.74**	**0.85**

Figure 4.2: Annotations for LA and NY. Top: percentages of user labels for the different categories. Bottom: absolute numbers of labeled users and annotation agreement results.

To explore different scenarios, we represented users as feature vectors, using three sets of features: **geographic** features, such as counts or percentages of visits to locations; **semantic** features derived from Foursquare, such as the popularity of visited venues or counts of visits to venues with certain categories like "Restaurant" or "Park". and **Census** derived features, such as the average ethnic makeup of all visited locations or the ethnic makeup of a user's most-visited location.

We performed all our experiments using the scikit-learn library [23] and tested the algorithms logistic regression, decision trees, naive Bayes, and support vector machines (SVMs). The results of our best-performing algorithms are displayed in Table 4.1

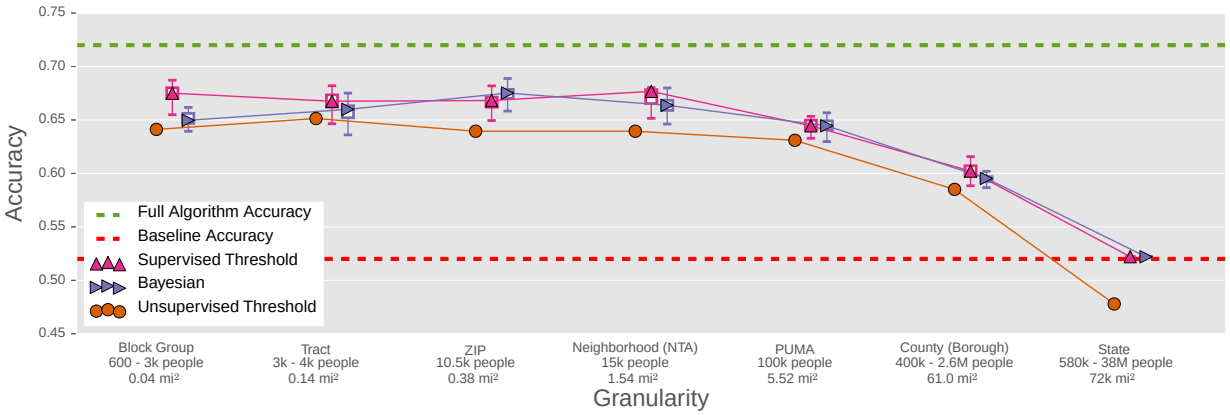


Figure 4.3: Accuracy of ethnicity prediction versus granularity for our NY population using several different inference techniques. Accuracy increases slightly at the ZIP code and neighborhood granularities and then decreases. Interestingly, the Bayesian algorithm, which uses only counts of visits to locations, performs comparably to the Supervised Threshold algorithm, which uses data on the ethnicity of visited locations.

Impact of Granularity and Feature Availability A detailed comparison of accuracy as a function of granularity and inference scenario can be seen in Figure 4.3. Granularity decreases (each location is a larger geographic area) as the X axis increases, and the Y axis corresponds to accuracy. The color of each line corresponds to a different inference scenario. The **Bayesian** algorithm uses just the counts of location visits as features. This corresponds to a scenario where labels are available but locations are anonymized. The **Unsupervised Threshold** algorithm uses *no* labels, but rather averages the percentage of Caucasians (or men/women) living all locations that a user visits, outputting a label based on if this average is below or above the city's mean. This corresponds to a problem scenario where locations are known but users are not labeled. In comparison, the **Supervised Threshold** algorithm learns a threshold via the labeled data, rather than using the city's mean, corresponding to a scenario where there are some labeled users and "raw" or lightly anonymized location data that can be linked to the census. The dashed lines in the graph correspond to our best performing algorithm and the baseline.

A few interesting results can be derived from Figure 4.3. Unsurprisingly, the performance of all algorithms decreases at the most coarse granularities. However, the performance stays remarkably flat up to fairly large granularities, without a significant drop until after the PUMA level of roughly 100,000 people. Additionally, the performance of the Unsupervised Threshold algorithm, which uses *no* labels, is not far from the other algorithms. Several algorithms improve in performance at medium granularities, such as ZIP and neighborhood, most likely due to the sparsity of our dataset at the most detailed granularity levels.

Impact of Data Quantity and Diversity Finally, with four different analyses, we studied the impact of data quantity on prediction accuracy. We examine algorithm accuracy on users grouped according to their number of geolocated Instagram photos and unique ZIP codes. Both of these are impacted by choices made by users—users who post more might be inherently easier to identify or predict. We thus did two more analyses where we sampled locations from a user's full set

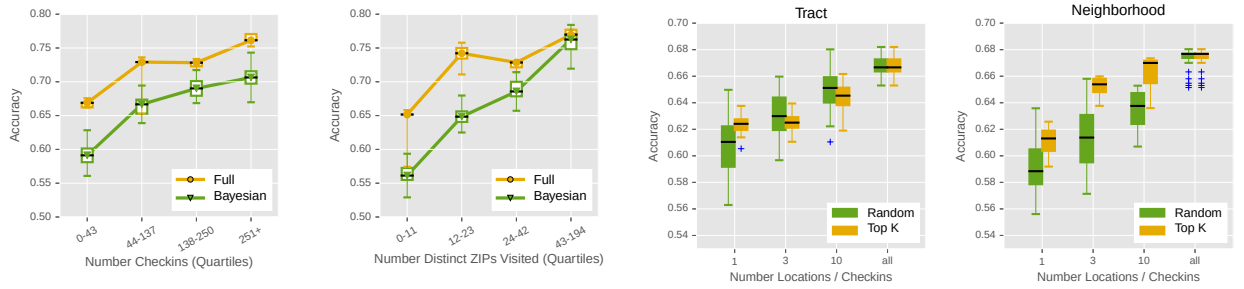


Figure 4.4: Impact of location quantity and diversity on accuracy. Y axis is accuracy, x axis is (from left to right): (1) Number of checkins used (2) Number of unique ZIP codes visited (3) Number randomly selected or Top-K Tracts (3) Number randomly selected or Top-K Neighborhoods

of checkins. In the first, we ran the Supervised Threshold algorithm on a user's k most visited locations. In the second, we ran the Supervised Threshold algorithm on n randomly sampled checkins. The results of this analysis can be seen in Figure 4.4.

4.2 Inferring Demographics from Social Media

In this section, we evaluate the feasibility of combining publicly available social media images and metadata (including location) with modern facial recognition techniques to conduct large scale demographic research. We find that facial recognition can be used to label the gender and race of social media profiles with high precision and recall. We further investigate factors that improve or hinder demographic labeling accuracy, showing a disparity between the accuracy of labelings of profiles of racial majority and minorities.

Name	*Users*	*Photos*	*Geotagged*	*Prosograms*
Full	260,389	115M	16.5M	
Face	4,166	1.5M	64.6k	322k
Labeled	172	16,655	5,489	5,272

Table 4.2: Overview of Datasets

For our research, we used a large dataset of Instagram photo metadata obtained through Instagram's API under the Terms of Service. Similar to the work described in Chapter 4.1, we gathered the metadata (such as time of photo, URL of image, tags, location, etc.) for all photographs of a "root" user, Kevin Systrom, the founder of Instagram, and then collected the user IDs of users who had commented on or "liked" his photos, gathered their metadata, and repeating this process in a random outward searching manner. This process resulted in over 115 million photos on over 260,000 profiles. A summary of this data (and derived data described in subsequent sections) is displayed in Table 4.2.

For subsets of profiles, we gathered additional information. For all photos, we included geographic information when available. For a randomly selected subset, we ran a publicly available

face recognition API[1] on a users first 100 Instagram images. In addition to recognizing faces within images, Face++ labels race from {White, Black, Asian} with a *confidence score* 0-100 and gender from {Male, Female} with a *confidence score* 0-100. For a smaller subset of users, two research assistants labeled a randomly selected subset for the collected profiles for gender and race, with 98.8% agreement on gender and 85.5% agreement on race, with the labels consistent with the Face++ set.

Name	*Cost*	*Speed*	*Accuracy*
Manual	Expensive	Moderate	High
Face Recognition	Free / Cheap	Slow	Moderate
Census	Free	Fast	Low

Table 4.3: Demographic Labeling Techniques

There are many different ways to go from the "raw" signal of social media to clean labels of demographics. In this section, we consider three, with the relative comparison of the costs and benefits displayed in Table 4.3. **Manually** labeling profiles using humans to examine each one will result in the highest accuracy, but will have high monetary costs. Time for labeling can be moderate if done in parallel with many labels, and much slower if done serially by few individuals.

Race	*Predicted Minority*	*Predicted White*
Minority	8	31
White	5	88

Table 4.4: Location-Based Race Labeling Confusion Matrix

Face recognition can provide a valuable signal for gender or race based on image analysis. Depending on the implementation, this can be very cheap, as it is bounded by computation. However, face recognition algorithms are imperfect and may exhibit statistical bias towards different groups, and may be slow if bounded by API services or computation. Finally, using **location data**, as described in Chapter 4.1, can be incredibly fast and cheap, but will have low accuracy depending on prediction task, location granularity, and other factors. We next explore in detail the accuracy of each of these methods. For the sake of brevity, these results include only information on predicting US Census Race classifications, but we have additional work on binary gender classification available as well.

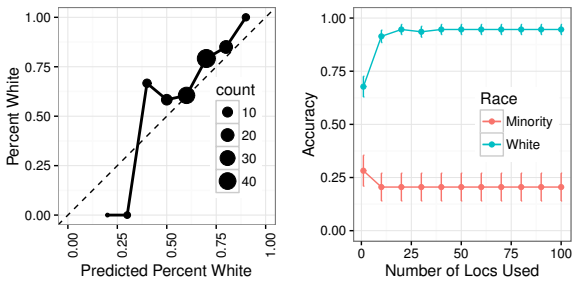


Figure 4.5: Location Prediction: Calibration and data amount to accuracy of race prediction.

Based on previous results, we expected location to provide more of a signal for race than for gender. Census tracts in the United States show a much more skewed distribution in regards to race than to gender, with the majority of tracts being heavily Caucasian, but with a sizable fraction of

[1] `faceplusplus.com`

tracts (>6%) having smaller than 20% Caucasian. On a baseline of 70.4% white users (including Hispanics), using the average percentage of census tracts visited predicted 73.4% of users' races. Our figures show that there is some weak signal in terms of ROC. The algorithm is well calibrated but again does not make strong predictions on most users. Accuracy is very poor on minorities, compared to Whites, and after an initial jump in accuracy after ten locations, is not greatly affected by the number of geotagged in the users profile. The number of locations used seems to have little effect on accuracy, which remains low for minorities and high for Whites.

Race	*Labeled Minority*	*Labeled White*
Minority	33	11
White	4	97

Table 4.5: Content-Based Race Labeling Confusion Matrix

In Fig. 4.6, we see that the algorithm is under-confident, outputting a lower probability than warranted when predicting if a user is white. An important aspect of demographic labeling is considering issues of the digital divide or disparate impact. In the rightmost plot of Fig. 4.6, we see that accuracy is much lower on minorities than it is on White users. Errors seem biased towards underrepresentation of minority groups. This could have consequences as the algorithms introduce a hegemonic factor in favor the majority.

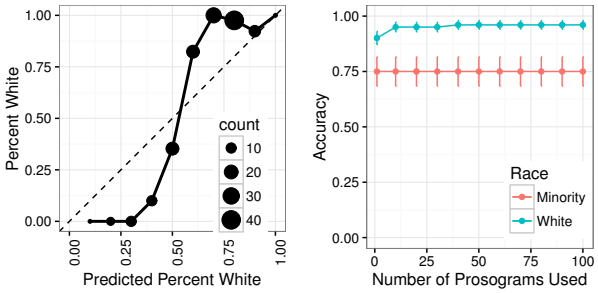


Figure 4.6: Face prediction: Calibration and data amount to accuracy of race prediction.

In both our face recognition and location-based labeling methods, as the number of photos used increases, the impact of one individual photo goes down. We therefore should expect that at some point we should observe "diminishing returns" in terms of the amount of data collected. In other words, we do not need to crawl all photos of a profile in order to know what our algorithm will label it. We investigate this idea in Figure 4.7.

We first gather all face data for each user, labeling their entire profile. Using all of the data, we a label a profile for gender and race using our algorithms. For each user, we then restrict the data input to the algorithm to a smaller subset and record our prediction. We repeat this process, gradually decreasing the amount of data supplied to the algorithm until we only give one photo. For each number of prosograms used, we calculated what percentage of users have the same label as their label from the full labeling. We see in the left figure that if we used just 1 picture, the output of our algorithm would agree with the full data about 75% of the time, and this is the same for both male and female profiles. After 100 prosograms, very few profiles will change their label. This suggests that for gender, we need only collect the first 100 prosograms to get a highly accurate estimate on our algorithm's result on the entire user profile data.

We observe similar behavior in the race dataset, depicted in the right side of Fig 4.7. After 100 photos, our algorithm is similarly unlikely to change its label. However, in contrast to gender, we see divergent levels of agreement for our two labels, with profiles eventually labeled minority

more likely to disagree than white profiles. This points to a potential bias in either the data or in the labeling algorithm, again suggesting that care must be taken in order to achieve balanced error rates across classifications.

Debiasing Oftentimes, scientists might choose to optimize accuracy or a class-centric metric like F1. However, choosing to optimize for metrics focused on one class (such as Female or White) can cause systemic biases to appear, as accuracy improves for one class but is ignored for another. This can be of particular in demographic labeling, where we would like to have proper representation of multiple groups. For example, when using the threshold of 0.5, we see large disparities between White and Minority labels as the data scales, as appears in Figures 4.5 and 4.6.

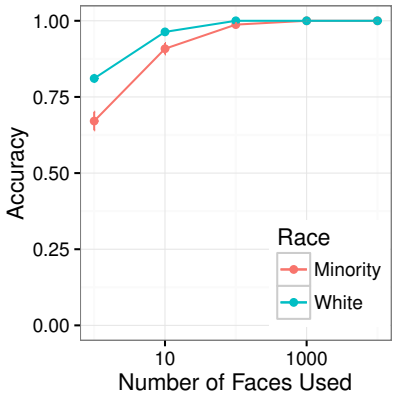


Figure 4.7: Agreement as data increases

Instead of optimizing for accuracy or F1, one method to insure equal levels of error in both classes is to instead optimize for Balanced Error Rate (BER). Balanced error rate is simply an average of the error in each class, that is, $\frac{1}{2}\epsilon_1 + \frac{1}{2}\epsilon_2$, where ϵ_1 and ϵ_2 represent the error in class 1 and 2 (e.g. Female and Male, or White and Minority).

Figure 4.8 shows how accuracy scales with more data for race when using a threshold determined by optimal BER. The thresholds were 0.58 for faces and 0.67 for location. Comparing with Figures 4.5 and 4.6, we see lines much closer together, indicating error rates much closer to one another than previously. Choosing another metric, such as F1, indeed leads to problems in the example of location. Here, a threshold of 0.33 is chosen. Although this gives an accuracy of 100% for White, it affords only 10.3% accuracy for minorities.

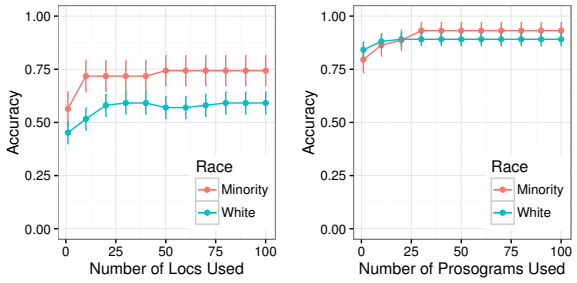


Figure 4.8: Location or faces vs. accuracy, with a choice of threshold determine by Balanced Error Rate.

Using BER as a metric in training can help us in designing algorithms that will fairly label users when conducting demographic research. Such techniques are especially important when being used in unsupervised settings, where algorithmic bias can impact accuracy on a large scale.

4.3 A Personal Location Data Auditing Tool

In this thesis, we show how demographics can be inferred from location data, but many users are not aware of the privacy implication of the collection of their information. This section, which appeared WWW in 2016 [24], shows a tool that we created in order to inform users, regardless of technical skill, about what their location information can reveal. We will begin with a summary of

a typical use of our tool “FindYou", and proceed to explain each component in more detail, along with the decision-making that influenced the design.

Site Summary

When opening the site, the user is greeted with a general description of the project. After clicking through this screen, the user has the option to import their data from three different web services or to manually import data by clicking visited locations on a map. Upon importing their data, users see the distribution of their visited locations of several different demographic traits, including race, income, age group, and parental status. Finally, at the bottom of the page, users have the ability to donate their data for further research.

Design Decisions

Why did we choose these sites? FindYou is currently able to import data from three popular online services or manually, by a user clicking on visited points on a map. The three sites we chose are Instagram, Twitter, and Foursquare. These sites were chosen because they are all popular but also present a diversity of behaviors and different levels of focus on location. We will discuss each of these sites in turn.

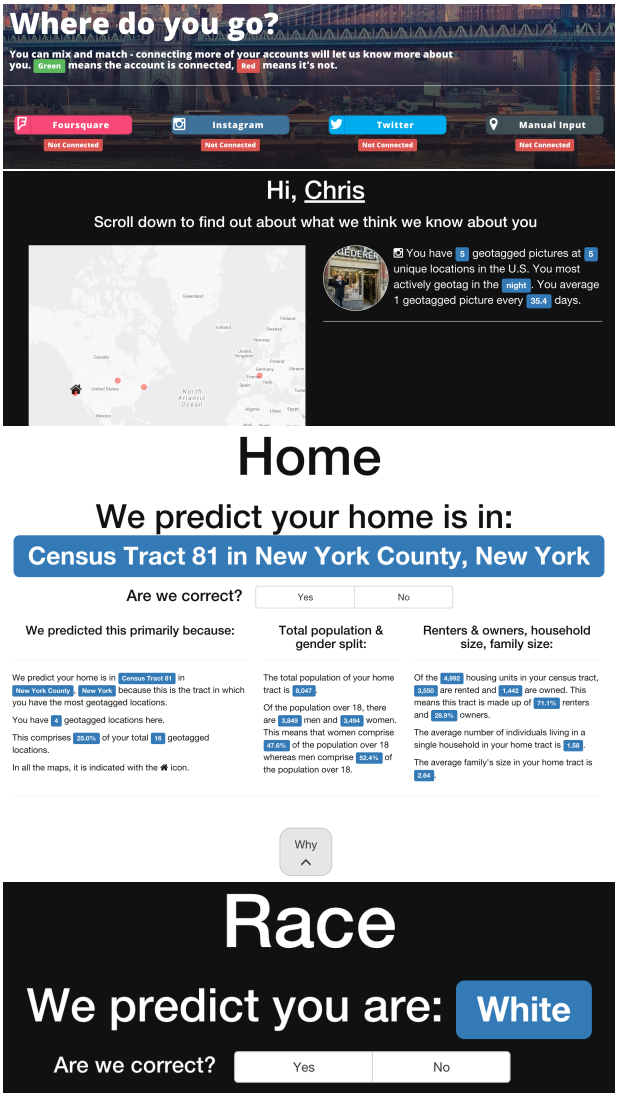


Figure 4.9: Screenshots from the site, displaying (top to bottom) options for linking to data sources, a map showing the users' data, and predictions about home location and demographic, with prediction details.

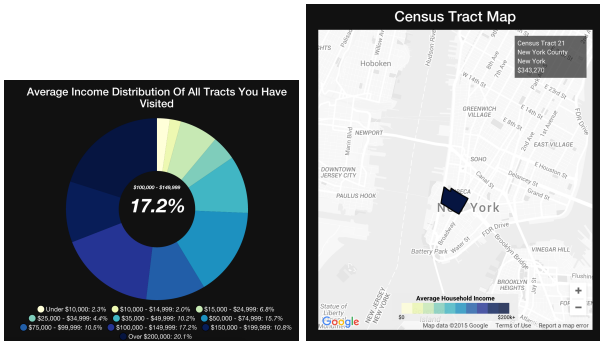
Foursquare is a location-based social network and review site. Users write reviews of and give tips about locations they have visited. It is estimated to have 50 million users. Foursquare is the most “location-centric" of our utilized web-services, as users must reveal their location to obtain any value from the service. **Instagram** is a photo-sharing application owned by Facebook with 400 million monthly active users. Instagram is notable for it being primarily targeted at mobile phones; currently users cannot upload photos from a desktop or laptop computer. The mobile focus makes it is easy for users to “tag" photos with locations using their phone's GPS device. Although many users do tag their photos with location data, unlike Foursquare, it is not necessary to post a location in order to use the app. Due to the fact that many users do tag their photos with locations, it is the second-most “location-centric" of our three services. **Twitter** is a microblogging service where users post 140 character texts called “tweets". Twitter has approximately 320 million users. Through its smartphone interface, Twitter users can tag tweets with locations. Many users connect their Twitter account to other web services, such as Foursquare and Instagram, among others, which may also contain location

data. The primary focus of most tweets is not about where a user currently is. Therefore, Twitter is the least location-centric.

We additionally included an option for **manual input**. This option simply has users click on a map to say where they've been. We included this option and used this design for several reasons. First, we wanted users who do not use any of the three aforementioned services to be able to participate in a location information privacy audit. Additionally, allowing users to manually input data gives the ability for users to play with hypothetical trips or to input locations that were not tagged in the services. We used this design because it is easy and simple.

Why did we choose to display these demographic features? After a user has imported at least some of their location data, we display demographic information on the places they visited. The features we chose to show are race, income level, age, and family make-up (number of households with children). The user sees a pie chart showing the average (over the user's visited locations) categorical distribution for that demographic trait. The site additionally displays specific details about each category for the user's most visited location. Technically, this works by utilizing information from the United States Census. On our server, we store information on the boundaries of each U.S. Census tract. We additionally have information on the make-up of each Census tract for our selected traits. We chose these features to be interesting, surprising, and possible to infer using location data. Hopefully, FindYou can include additional interesting demographic features in the future.

Why did we use only simple machine learning techniques? In addition to descriptive data about the distribution of visits in each category, we also present predictions of which category a user falls into for each demographic attribute. Although users may be interested about the demographics of the locations they visit, they might not realize that this information can be used to infer their own traits. Therefore, showing predictions is useful in and of itself, even if the predictions aren't all accurate, as it shows users that their data can be used in such inferences. Driven by our goal of simplicity in explaining what's going on to the user, we use simple techniques that are intuitive for most users, as opposed to using more difficult to understand methods like SVMs or neural networks. For each demographic trait, we predict the user to be in the class to which they have the most visits. To make this concrete, consider the example of age. We break age into several categories. We average the distribution of age categories of all the locations a user has visited, and pick the category with the largest proportion.



(a) (b)

Figure 4.10: (a) Donut graph displaying distribution of income groups visited by user, and (b) map showing tracts visited by user along with income information on each tract.

How did you choose to represent locations? There are many different ways to represent locations, such as latitude longitudes, venues, cities, or points of interest. Throughout the paper and the site, we use a United States Census tract as an "atomic" location. The United States Census partitions the country into *census tracts*, which are stable geographic boundaries chosen to contain homogeneous populations. Census tracts are typically the size of a few city blocks and might con-

tain 4000 or fewer people. We chose to represent all locations as a census tract for several reasons. First, we can map any latitude longitude point into a census tract, and thus any venue with an associated lat-lon into a tract as well. Census tracts are small enough to be targeted, but large enough to display without overwhelming the user. Finally, they are all associated with detailed demographic information from the Census. Throughout the site, whenever a census tract is mentioned, the user can click on it to see its geographic boundaries and demographic make-up.

Chapter 5

Proposal

As described in Chapter 4, an important challenge facing the computer science community is algorithmic bias. In recent years, an emerging body of work has focused on different mitigating techniques, such as automated discovery of bias, “de-biasing" existing algorithms, or theoretical analyses of different types of bias. De-biasing techniques are sure to incur a cost: the objective function of the algorithm is no longer as straightforward, and organizationally new infrastructure needs to be put into place for something that could hurt revenue. Understanding the key trade-offs between revenue and uncertain risk will be important to insure real-world adoption. Although there have been some good initial insights, the community has lacked strong data-driven analysis on this trade off.

I propose to fill this gap by applying proposed techniques to real-world problems through the use of an innovative dataset. Namely, I will look at the real-world problems of recommendation systems within a large social network, focusing on location-based advertising. As locations are an important signal in user intent and demographics, they are a useful system for advertising. By the same token, since locations are tied to demographics, an automated advertising system based on advertising may become biased against certain groups (inadvertently or otherwise). An example unfair outcome is if certain racial groups or genders are not offered deals or benefits at the same rate as others due simply to their location.

5.1 Background

Algorithmic bias has been studied in many contexts. The common idea is that algorithms are not completely neutral and objective arbiters of decisions, but rather are simple tools, the incautious use of which can lead to harm falling disproportionately on already vulnerable groups. A prototypical example was the discovery of a chain that altered its online prices based on the buyer’s home locations, possibly in an attempt to give discounts to users living near competitors. However, living near competitors correlates with income, in effect raising prices for lower income buyers [34, 21]. Work in the popular press has found other cases, such as differing types of errors across race in algorithms designed to help with bail sentencing [2] and higher car insurance rates for minorities compared to whites living in areas with the same level of risk [1]. Other cases of bias have been caused by training algorithms on data that already contain human biases, for example word embeddings that associate doctors with men and women with nurses [4, 6]. Researchers have

attempted to mitigate these issues by finding and subtracting dimensions that correspond to gender.

A variety of methods have been proposed to automatically remove bias from machine learning algorithms. One example attempts to learn the protected class of individuals in parallel to the original machine learning task, altering the data to be less distinguishable to the demographic differentiator while still minimizing the original loss function [11]

My plan is to utilize the algorithm in "Fairness Through Awareness" [10]. The proposal of this algorithm is that some fair arbiter decides on a similarity scoring function which will be applied between users. It adds constraints to an algorithm that users similar on the similarity scoring function metric must have similar outcomes. The scheme creates a linear program with constraints polynomial to the number of users and outcomes, making it practical to implement. Indeed, I have already run this scheme on toy examples and do not anticipate large difficulties on more complicated versions of the problem.

More rigorously, "Fairness Through Awareness" has the algorithm designer solve the following linear program:

$$\begin{aligned} \min \quad & \mathbf{E}_{x \sim V} \mathbf{E}_{a \sim \mu_x} L(x, a) \\ \text{subject to} \quad & \forall x, y \in V : D(\mu_x, \mu_y) \leq d(x, y) \\ & \forall x \in V : \mu_x \in \Delta(A) \end{aligned}$$

where

- V is the set of all representations of users
- A is the set of outcomes (e.g. the ads to show users)
- μ_x a distribution for user x over the outcomes A
- L is the loss function we are trying to minimize (such as negative revenue)
- D is a distance function between distributions, such as total variation or l_{inf}
- d is a distance function between representations of users

Intuitively, this linear program tries to minimize loss while showing assigning a similar distribution of potential ads to show to similar users. I believe this framework allows for a high level of flexibility for the algorithm designer while still using an intuitive and effective notion of fairness.

5.2 Research Plan

Much of the previous work has either focused on detecting bias. Researchers build tools for auditing advertisers, pointing out cases where algorithms have lead to undesirable results. Additionally, many works have used small and limited datasets or included no data-driven analysis at all. I propose to fill this gap by applying proposed techniques to real-world problems through the use of an large and innovative dataset, focusing on a practical and important application: location-based advertising. As locations are an important signal in user intent and demographics, they are a useful system for advertising. By the same token, since locations are tied to demographics, an automated

advertising system based on advertising may become biased against certain groups (inadvertently or otherwise). An example unfair outcome is if certain racial groups or genders are not deals or benefits at the same rate as others due simply to their location.

At a high level, I plan to first model a location-based advertising system using a large amount of metadata collected from a location-based social network. This data will include user profiles of visited locations, as well as social network information such as text and Demographics of profiles will be determined based on the face recognition system described in Chapter 4.2. We will simulate advertising by predicting when users take a particular action, using that as a proxy for a click, associating a revenue with that action commensurate with a cost-per-click (CPC). This model will give us an objective function which we can then try to maximize in the setting of Dwork's de-biasing algorithm. We can use a variety of similarity functions, such as earth mover distance between the histogram of locations visited by a user. This will let us see when statistical parity can be achieved between various demographic groups based on their mobility. We can additionally experiment with other similarity functions, looking for the interaction between revenue loss and increased fairness.

The project will be conducted in several steps.

Task	Status	Time line
Collection of data.	Completed	Aug '16
Labeling of data with Face++ API.	Completed	Aug '16
Initial analysis and descriptive statistics of dataset	In progress	Start May '17
Full problem specification: algorithms, inputs, and objectives.	In progress	Mid May '17
Apply de-biasing to algorithms and analyze impacts.	To do	Early June '17
Create recommendations for algorithm designers.	To do	Mid June '17

Table 5.1: Plan for completion of my research

Data. I plan to use a dataset collected from the popular image sharing service Instagram. Currently, I have metadata for over 115M public photos for 260,000 users, collected via Instagram's API. 16 million of these photos have location data, comprising 162,000 users from 180 different countries. I have applied Face++'s face recognition API on 2 million photos for over 6000 users, obtaining gender and ethnicity data on 844,000 faces. Based on the labeling schemes described in my previous work, I believe 3,375 of these users are female and 2,859 are male, with 5,027 Caucasian and 1,207 non-Caucasian.

Problem description. I have several different ideas for problem scenarios. One is *tag prediction*. Instagram users put "tags" on their photos to indicate the category. These tags could be used to indicate interest– for example, a user that may be interested in a coupon for a fast food restaurant may tag a photo "#burgers". I plan to categorize tags based on some affiliation with demographics and with locations. The model advertising scenario will be a predictive model of tags (or category of tags) based on mobile behaviors. Given the time, location, and user profile of a photo, we will predict what tag a user might use. If a user does indeed use that tag in that photo, we will record that as a click on an a. We can associate tags to a monetary value by looking at current CPCs on Google AdWords or another online advertising service.

Another idea problem scenario could be *trip prediction*. An advertiser may wish to target users who are likely to buy a plane ticket or other expensive travel-related item. There may be some signal in previous trips or checkins at airports, hotels, or other travel-related locations. The goal is to predict who will travel in a future time period based on current behaviors. The idea is then to de-bias this prediction scheme to insure fairness in showing deals across demographic groups.

User representations and similarity measures. A nice property of the earth mover distance as a user distance measure, is that if the user distance score is less than 1 for all pairs, a measure of bias between two groups will in fact be equal to the earth mover distance between the expected representation of two users sampled from each demographic group. Other user similarity/distance functions can also be applied, and I will investigate several distribution distance functions.

5.3 Current Results

The next major step in this work is developing the computational advertising model. As mentioned above, I am proposing two different formulations of the model: (1) predicting tags and (2) predicting large (out of state or country) trips. Due to the massive number of tags (over 645 million tags on the 115 million photos), it will be necessary to find a subset of tags that are predictable from location and have a high potential to be shown to one demographic but not another.

Demographic-affiliated tags. Utilizing my set of users labeled with demographics from the face recognition, I was able to associate some tags with each demographic. I then used a χ^2 test to find tags most associated with one demographic (using a one vs many approach). This yielded the following results:

Tags associated with female profiles: "makeup", "nails", "dress", "lipstick", "sisters", "pink", "girls", "nailpolish", "nailart", "lips", "bikini", "brunette", "ootd", "yummy", "hair", "nail", "bff", "mylove", "chocolate", "girly".

Tags associated with male profiles: "beard", "hiphop", "gay", "dj", "muscle", "guy", "skate", "dope", "man", "brasil", "vinyl", "brazil", "cars", "bike", "building", "jj", "tour", "graffiti", "brooklyn", "urban", "instagay", "gayboy", "bodybuilder".

Tags associated with Asian profiles: "indonesia", "japan", "singapore", "korea", "thailand", "tokyo", "bangkok", "bali", "asian", "hongkong", "japanese", "beach", "sea", "cafe", "dog", "cat", "sunset", "coffee", "sky","flower"

Tags associated with Black profiles: "wcw", "tbt", "truth", "hiphop", "dope", "blessed", "god", "repost", "motivation", "nofilter", "goals", "faith", "beautiful", "inspiration", "art", "nyc", "friends", "picoftheday", "fall", "eyes"

Tags associated with White profiles: "summer", "winter", "sun", "spring", "autumn", "snow", "clouds", "flowers", "italy", "nature", "mountains", "relax", "blonde", "puppy", "streetart", "water", "lake", "blackandwhite", "summertime", "sunglasses"

Location-affiliated tags. The next step in the process is to find tags that are likely to be predictable from location. A variety of statistical measures have been proposed to understand the relationship between some variable and geographic. Specifically, spatial autocorrelation is typically used, with a measure such as Moran's I or Geary's C. A key issue that we encountered is a level of granularity. In the United States, populations (especially the Instagram using population) is clustered in few areas, areas with large cities. Thus, it is more important to find areas with a high concentration of tags, as opposed to finding tags that are in neighboring geographies. Therefore the simple measure of L_2 norm may be a suitable choice.

Next steps. The next steps in this project will be to find the most useful representation for location of users, create the advertising model (and evaluate its efficacy), and inspect the relationship between fairness and revenue.

Bibliography

[1] Julia Angwin, Jeff Larson, Lauren Kirchner, and Surya Mattu. Minority Neighborhoods Pay Higher Car Insurance Premiums Than White Areas With the Same Risk. *ProPublica*, 2017.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias: ThereâĂŹs software used across the country to predict future criminals. And itâĂŹs biased against blacks. *ProPublica*, 3, 2016.

[3] Jon Bing. Classification of personal information with respect to the sensitivity aspect. *Databanks and Society*, pages 98–150, 1972.

[4] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357, 2016.

[5] Laura Brandimarte et al. Misplaced confidences: Privacy and the control paradox. *WEIS*, 2010.

[6] Aylin Caliskan-Islam, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora necessarily contain human biases. *arXiv preprint arXiv:1608.07187*, 2016.

[7] Juan Pablo Carrascal, Chris Riederer, Vijay Erramilli, Mauro Cherubini, and Rodrigo de Oliveira. Your Browsing Behavior for a Big Mac: Economics of Personal Information Online. In *WWW '13*.

[8] Yves-Alexandre de Montjoye et al. Unique in the crowd: The privacy bounds of human mobility. *Sci. Rep.*, 3, 2013.

[9] 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.

[10] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. ACM, 2012.

[11] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.

[12] Saikat Guha et al. Koi: A Location-Privacy Platform for Smartphone Apps. In *USENIX NSDI*, 2012.

[13] Saikat Guha, Mudit Jain, and Venkata N Padmanabhan. Koi: a location-privacy platform for smartphone apps. In *NSDI'12: Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*. USENIX Association, April 2012.

[14] Sibren Isaacman, R Becker, R Cáceres, S Kobourov, Margaret Martonosi, J. Rowland, and A. Varshavsky. Identifying important places in people's lives from cellular network data. *Pervasive Computing*, pages 133–151, 2011.

[15] Sibren Isaacman, R Becker, R Cáceres, S Kobourov, Margaret Martonosi, J. Rowland, and A. Varshavsky. Ranges of human mobility in Los Angeles and New York. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on*, pages 88–93, 2011.

[16] Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, James Rowland, and Alexander Varshavsky. A tale of two cities. In *HotMobile '10: Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications*. ACM Request Permissions, February 2010.

[17] Patrick Gage Kelley et al. When are users comfortable sharing locations with advertisers? In *ACM CHI*, 2011.

[18] K. Krippendorff. *Content analysis: An introduction to its methodology*. SAGE, Beverly Hills, CA, USA, 1980.

[19] John Krumm. A survey of computational location privacy. *Personal and Ubiquitous Computing*, 13(6):391–399, 2009.

[20] Ilias Leontiadis et al. Don't kill my ads!: balancing privacy in an ad-supported mobile application market. In *ACM HotMobile*, 2012.

[21] Jakub Mikians, László Gyarmati, Vijay Erramilli, and Nikolaos Laoutaris. Detecting price and search discrimination on the internet. In *HotNets-XI: Proceedings of the 11th ACM Workshop on Hot Topics in Networks*. ACM Request Permissions, October 2012.

[22] A Narayanan and V Shmatikov. Robust De-anonymization of Large Sparse Datasets. *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pages 111–125, 2008.

[23] F. Pedregosa et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[24] Christopher Riederer, Daniel Echickson, Stephanie Huang, and Augustin Chaintreau. Findyou: A personal location privacy auditing tool. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 243–246. International World Wide Web Conferences Steering Committee, 2016.

[25] Christopher Riederer, Vijay Erramilli, Augustin Chaintreau, Balachander Krishnamurthy, and Pablo Rodriguez. For sale : your data: by : you. In *HotNets-X: Proceedings of the 10th ACM Workshop on Hot Topics in Networks*. ACM Request Permissions, November 2011.

[26] Christopher Riederer, Yunsung Kim, Augustin Chaintreau, Nitish Korula, and Silvio Lattanzi. Linking users across domains with location data: Theory and validation. In *Proceedings of the 25th International Conference on World Wide Web*, pages 707–719. International World Wide Web Conferences Steering Committee, 2016.

[27] Christopher Riederer, Sebastian Zimmeck, Coralie Phanord, Augustin Chaintreau, and Steven M Bellovin. âĂIJi donâĂŹt have a photograph, but you can have my footprints."–revealing the demographics of location data. In *ACM Conference on Social Networks*, 2015.

[28] Christopher J Riederer, Augustin Chaintreau, Jacob Cahan, and Vijay Erramilli. Challenges of keyword-based location disclosure. In *WPES '13: Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society*, pages 273–278, New York, New York, USA, November 2013. ACM Request Permissions.

[29] Luca Rossi and Mirco Musolesi. It's the Way you Check-in: Identifying Users in Location-Based Social Networks. *COSN '14: Proceedings of the 2nd ACM conference on Online social networks*, pages 1–11, August 2014.

[30] United States Census Bureau. 2010 census. http://factfinder2.census.gov/faces/nav/jsf/pages/index.xhtml, 2010.

[31] United States v. Jones. 2012. 132 S. Ct. 945, 955 (Sotomayor, J., concurring) (quoting People v. Weaver, 12 N.Y.3d 433, 441-42 (2009)).

[32] J Unnikrishnan and F M Naini. De-anonymizing private data by matching statistics. In *Communication, Control, and Computing (Allerton), 2013 51st Annual Allerton Conference on*, pages 1616–1623, 2013.

[33] Jennifer Valentino-Devries and Jeremy Singer-Vine. Websites vary prices, deals based on users' information. *Wall Street Journal*, December 24 2012.

[34] Jennifer Valentino-Devries, Jeremy Singer-Vine, and Ashkan Soltani. Websites Vary Prices, Deals Based on Users' Information. *online.wsj.com*, pages 1–6, December 2012.

[35] Hui Zang and Jean Bolot. Anonymization of location data does not work: a large-scale measurement study. In *MobiCom '11: Proceedings of the 17th annual international conference on Mobile computing and networking*. ACM Request Permissions, September 2011.

[36] Jiawei Zhang, Xiangnan Kong, and Philip S Yu. Transferring heterogeneous links across location-based social networks. In *WSDM '14: Proceedings of the 7th ACM international conference on Web search and data mining*, pages 303–312. ACM Request Permissions, 2014.