Big Location Data:
Balancing Profits, Promise, and Perils

Thesis Proposal

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

March 28, 2017

Abstract

The “Big Data" era has a lot of potential. Businesses hope that Big Data will make them more efficitient and profitable or enable entirely new products. Governments hope to provide better for the needs of their citizens. At the same time, these new collections of data can present societal risks, as we've now enabled mass surveillance, a loss of privacy, and algorithmic bias.

The rise of cheap and ubiquitous electronics, including but limited to smartphones, has enabled the capture and use of human mobility data like never before. As a subset of Big Data, 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, or more private characteristics of race, religion, sexuality, or political affiliation.

In this thesis proposal, I describe recent work that attempts to balance the scientific and engineering promises of location data with the potential risks. I will describe work I have completed relating location data to privacy, anonymity, economics, and algorithmic bias. I propose future research to be completed in the form of a thesis, advancing knowledge of location-based demographics and algorithmic bias.

Contents

1 Introduction

In the United States v. Jones decision, 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" [74].

1.1 Outline

I will begin with a background section which introduces the core concepts found in this proposal: location data, privacy, and bias. I proceed with three chapters detailing completed work. Each chapter contains a section summarizing relevant prior work.

Chapter 3 focuses on location data, privacy, and economics. 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.

Chapter 10 examines the possibility of anonymizing location data. Prior work has shown that users are highly unique in their location patterns, leaving them vulnerable to deanonymization (see 2.2). 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.

Chapter 11 shows the potential for location data to be part of systems that 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 debiasing of algorithms.

Chapter 12

Chapter 13

I conclude with a plan for completing this work in Chapter 14.

2 Background

2.1 Location Data

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 unqiuely 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[1], 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 [2].

What is location used for? TODO

2.2 Privacy

Privacy has been an important concept, brought to the forefront of public debate as surveillance of users has grown, both by governments and private companies.

ADD MORE STUFF

In this work, we will focus on two technical conceptions of privacy, *k-anonymity* and *differential privacy*.

[1] `http://www.mapmyrun.com/`

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

k-anonymity
Differential privacy

2.3 Bias

2.4 Online Advertising

3 Location Data, Privacy, and Economics

The online economy is based primarily on advertising. The income of a firm roughly translates to (number of impressions) x (dollars per impression). I am trying to keep this abstract and not saying that firms are always getting paid for impressions, as other models like paying per click or per sale or other action are quite common. Really the argument here is that firms make money 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. TODO: cite some stuff.

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 utilty 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.

3.1 Related Work

3.2 Completed Work

3.2.1 Your Browsing Behavior for a Big Mac

How does one determine how a study participant values something as abstract as User privacy is extremely important. However, there does not exist a strong understanding of how users value their privacy.

3.2.2 For Sale: Your Data. By: You

[62]

3.2.3 Challenges of Keyword-Based Location Disclosure

[65]

4 Introduction

The rapid adoption of smart phones and tablets has led to innovative applications and services that exploit location information. Location information is increasingly used to drive advertising – location-based targeting generates four times as much revenue per impression compared to ads without location data[3]. Even brick-and-mortar stores use location data, with retailers using cell phones' WiFi signals to learn where customers spend time in their stores[4].

There are many privacy concerns surrounding the use of this data. For example, many applications access location information even when such information is not needed, and may share it with multiple third parties, leading to privacy concerns [17, 75] and attracting the attention of regulators [20, 74]. This work focuses on location information generated in real-time by users with mobile devices.

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 [42]. Because it affects their profits, companies that are a part of the mobile application ecosystem oppose any regulation that may restrict access to location data and claim that the "cost" of a privacy bill threatens the web's general economy and ultimately hurts customers. 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 contrast to previous work, we aim 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] `http://bit.ly/vXWdsw`
[4] `http://nyti.ms/15vLRva`

Our main contributions are: (1) The design of a keyword-based system that integrates well into today's location collection and monetization. Our solution requires no change on users' devices, a minimum level of indirection, and addresses goals like usability, deployability and scaling (Sec. 5). (2) A test of our solution's usability and relevance with a small scale trial on real users. While this experiment is too small to form statistically significant conclusions, it allowed us to test the feasibility of our design (Sec. 6). (3) An analysis of how such a system can offer different levels of protection against various threats, including freeriding, inference attacks using auxiliary information, and user misconduct (Sec. 7).

5 Overview

This section presents the motivation, design and advantages of a location disclosure system based on keywords.

5.1 A keyword-based solution

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 [25],

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 [36] 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 [25]. (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 [6]. 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 design we next describe is meant to operate under the following set of **assumptions**. Given the amount of press on privacy related issues, we believe that the PR backlash in the case of a se-

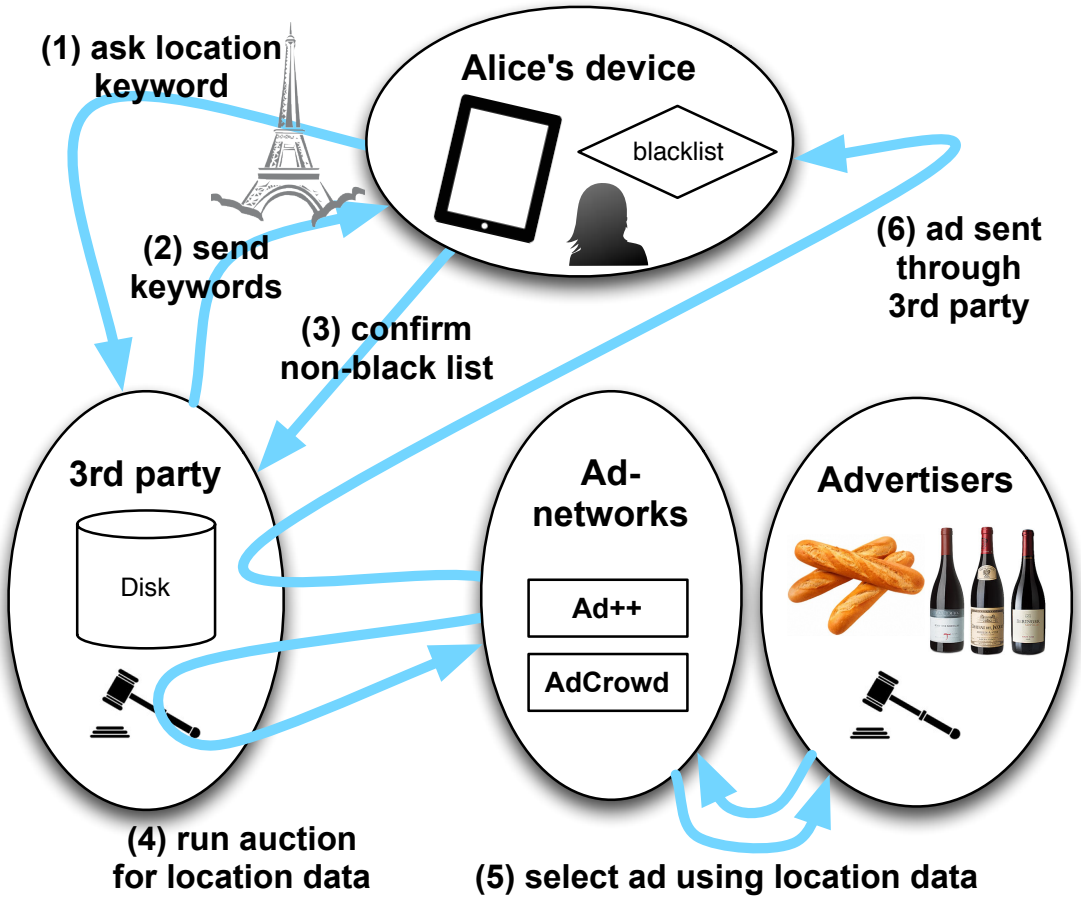


Figure 1: Solution overview

rious privacy violation will make such violations undesirable. As a consequence, we provision against an *honest-but-curious* advertiser. It means the adversary complies with the system but it can exploit the information that is gathered for its own interest. We provide safeguards against inference and linkage attacks. We also assume that the mobile OS used complies with user's privacy, hence not sharing location information with any application once the user stated that request. Note that the architecture presented next is oblivious to a background service model (passive, potentially continuous tracking) or a check-in model.

5.2 Design and Example

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.

A high-level diagram is shown in Fig. 8. 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 [36]. 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 describe in Sec. 6.1 how the blacklist formation can be simplified through nested menus and re-ordering. We assume the third party is trusted and leave lowering this requirement to future work.

As Alice arrives at the bakery, her network activity goes through the *blocking module* that runs a mix-network to conceal her real network address, and provides privacy protection like dropping cookies to third parties, overwriting `referer` headers etc. [40] (see Sec. 6.1 for more on implementation). At every location, Alice's device contacts the *keyword server* which translates locations to keywords. A check is then made against the blacklist to verify if Alice is comfortable releasing this information. If a location has multiple keywords and *any* of them are on the blacklist, it is considered private. Once a location passes the check, it is put on the *market* for sale with a unique user-id and the keywords. This user-id is generated independently and can be periodically changed. The information then is (UID_{Alice}, (lat_1, long_1), Bakery). As she arrives at the wine shop, the information on the market will be (UID_{Alice}, (lat_2, long_2), Wine Shop), as the wine shop also passes the blacklist test. Ad-networks can pay to *access* Alice based on these two locations released. The payment will be credited to Alice, with a small fraction taken by the third party. The third party then fixes a network address to reach Alice at the wine shop and conveys it to the ad-networks. Alice can receive a targeted ad (via an app or via SMS) for a particular wine.

As soon as Alice moves out of the wine shop, her network address changes and her location again is not known to anyone but the trusted third party. When she is close to the Libertarian party headquarters, the check against the blacklist returns a positive result, and this location is not revealed to anyone.

5.3 Summary of Advantages

Now that we've described the system, we discuss the benefits of the system for various parties.

Users obtain monetary payment for their data and privacy through choice. The architecture operates in the network and hence, users do not need to make changes to their devices. If information is leaked or shared between colluding ad-networks, these parties would have to gain access to the user to monetize this information – and unless these parties have paid, they are prevented from gaining access to the user. Hence, we protect against adversaries aiming to extract economic gain. We deal with adversaries who try to infer the identity of users or blacklisted keywords in Sec. 7.

The keyword system also benefits the user. If a user is visiting a place they are unfamiliar with, they may not be accustomed to what areas are privacy sensitive. Because keyword mappings work in any location, a user's privacy is protected even in unfamiliar areas. Additionally, a user may simply not realize the privacy sensitive nature of a location they are in. Because all traffic is directed through our system, if a user starts using a location-based service at a location they don't realize is privacy sensitive, our system can catch it and warn the user before they complete the action.

Ad-networks and aggregators can obtain non obfuscated data in a legal way, minimizing data breaches. As the data is 'bought', the ad-networks can micro-target. Ad-networks and advertisers

can easily make sense of the location data, as keywords are already used for context in current online advertising systems. Rather than having advertisers need to bid specifically for each location, ad-networks can simply run auctions for ad impressions in locations associated with specific keywords.

Application developers do not need to alter their code as we operate directly in the network. Applications serve as a conduit to show ads to the users, much as they do today.

Finally, mapping locations to keywords helps our system evaluation. Ad-networks constantly run many auctions of impressions to a customer searching for a specific term. Cost-per-click (CPC) data from ad-networks hence reflects the overall advertising demand on this topic. We show how CPC data may be collected and used to understand the economic value of locations.

6 Deployment and User Study

We now describe in detail how such a system could be implemented. We additionally discuss a small-scale deployment and user study we ran in order to demonstrate the system's feasibility.

6.1 Implementation

An implementation consists of the five components described in section 5.2: 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. This is a possible area for improvement; in future work, the radius of a query could change depending on an estimate of the device's current accuracy or a user's privacy preferences. The categories were then sent to the device.

Future implementations could likewise map locations to keywords by reusing online services such as Yelp, Google Places, and Foursquare. A "folksonomy" approach can be used where users label a map over time, possibly receiving incentive. To encourage tagging of privacy-sensitive locations, the system can allow anonymous tagging.

The **location blacklist** module was written as an Android application, using the phone's GPS. The app, available on Google Play[5], 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. Thus, a user could select and de-select whole categories of keywords with a single button press, but could also expand categories to select specific words. We placed categories previously defined to be sensitive [5] 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.

[5]Link to app: `http://bit.ly/13qOMqC`

For the purposes of our small scale user study, we did not create a **blocking module**. In a full implementation, it would be necessary to block any third-party advertisers who did not participate in the system. The connections to ad-networks and aggregators (AdMob, Flurry Analytics etc.) can be blocked by a proxy and spoofing the MAC address. All necessary proxies already exist: Privoxy comes with advanced filtering capabilities and handles rewrites of the HTTP headers like the 'referrer' header to prevent leakages of any form, and mitmproxy can handle SSL[6]. In addition, users could upload their SSH certificates to enable the module in the middle to masquerade as the user. From an application's perspective, no logic is broken. Even for location based services like Foursquare or maps, an unintentional checkin or a search at a private location can be prevented by checking against the blacklist – an added benefit.

As this deployment was meant for exploratory purposes, we did not 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. All participants received a $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 simulate the costs of disclosing information, we publicly displayed a user's non-blacklisted locations on a web interface, viewable at `keyword.cs.columbia.edu`. In a real system, a user would risk that her information is used improperly or released to those who might use it in a damaging way. We believed that publicly displaying a user's information simulated this risk. To increase the publicity of their information, we instructed users to post the link on a social media site, such as Facebook or Twitter, and email us a screenshot.

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

6.2 Deployment and Observations

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

[6] `www.privoxy.org`, `www.mitmproxy.org`

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.

7 Mitigating Attacks

Having introduced the design of the system, we now turn our focus to one of our key goals: protecting the privacy and value of system participants.

7.1 Attacks on the Value of User Data

Our system prevents an adversary from economically benefiting by using information about a user without properly compensating her.

Ad-networks may try to build up interest profiles of users over time in order to better target ads later *without* compensating the user. Even if a user's anonymous ID is changed regularly, human mobility patterns are periodic and somewhat predictable, making it easy to link a current anonymous ID to an older one[7]. Our system does not prevent such profiling, and it even makes it easier as the market announces which data is for sale. However, we ensure that this strategy has no economic benefit, for the following reason: all traffic flows go through a proxy, and an ad network who does not pay will receive the identity and location of a user, but a random temporary ID. Then the ad-network, although it has a rich profile of user u, is not able to recognize u as the recipient of an ad. For the same reason, ad-networks do not gain by colluding or reselling the information. Unless a payment is made, the identity and location of u is unknown, and the profile alone does not aid targeting.

A related issue is trajectory-based profiling. If an ad-network learns the habits of a particular user over time, the ad-network can show ads based on where a user *is likely to be* rather than paying for an exact location. Again, ad-networks must always pay to be able to access a user's identity. Care must be taken, however, to make sure that a user does not unwittingly display information about a visited blacklisted location based on her trajectory: *e.g.* location B is sensitive and locations A and C are not, and the only way to get to C from A is via B). If Alice checks in at point A and then at point C, ad-networks may infer that she visited B. Such attacks are not likely, and can be dealt with by ensuring that after visiting a blacklisted location a minimum amount of time has passed before disclosing a location.

One concern is if an app works to circumvent the proxies and leak information about either the location or the identity of the user. Against location leakage, one solution is to substitute a fake location to the app if it does not disrupt service [28]. An adversarial app could monitor the location market and try to associate an anonymous user profile with a particular device. Combined with a

[7]Note this profiling works on *non*-blacklisted locations only.

profiling attack, it can then send targeted advertisements without compensation by recognizing this device from now on. This is a costly attack and can be prevented if OSes separate their advertising services from applications [42] or if the users does not need a permanent ID for this application. Note also that, since UIDs are changed periodically, the profile cannot be updated without paying and hence loses some value over time.

7.2 Attacks on User Privacy

We study the robustness of our solution against a form of attack based on *inference*. We consider a malicious adversary whose goal is to predict the visits to blacklisted locations of a specific user with some accuracy. This may seem a priori impossible since whenever a user visits a blacklisted location, no information about this visit is sent or shared anywhere.

However, because mobility patterns tend to be periodic and similar people may have similar mobility patterns, an adversary may be able to discover something about a specific user's blacklist by comparing their publicly available location information with the full (including blacklisted) location information of 'compromised users'. This auxiliary location information could be obtained via hacking or a malicious or buggy application. Inspired by de-anonymization techniques based on auxiliary information [53], we now pose the following question: "Can an adversary with the full knowledge of the location information of a significant fraction of users predict the blacklisted locations of other users with high accuracy?" We test this on a large dataset of Foursquare checkins. Intuitively, the sparsity of locations and checkins in this dataset allows for strong attacks of this kind.

As in the de-anonymization technique, we consider a similarity score $\text{Sim}(u, v)$ between two users based on common visits. Let L_u denotes the places that are visited at least 1 time by u. We define similarity as:

$$\text{Sim}(u,v) = \sum_{l \in \mathcal{L}} \frac{1}{\text{span}(l)} \mathbb{I}_{l \in L_u \cap L_v} \text{ , for span}(l) = \sum_{u \in \mathcal{U}} \mathbb{I}_{\{l \in L_u\}} .$$

Note that by doing so we weight more the co-occurrence of a rare location as a sign of similarity between two nodes.

The attack then proceeds as follows. For a given keyword k, the attacker looks at all accounts that visited a location tagged with k. For simplicity we will say that such a user visits keyword k. These are the probes used to find similar users who are more likely to behave like them. For a given user u, the adversary first locates the $n = 10$ closest users that are compromised in terms of similarity $v_1, \cdots, v_n$. The attacker then computes the following weighted sum:

$$P(u) = \frac{1}{\sum_{i=1}^{n} \text{Sim}(u, v_i)} \sum_{i=1}^{n} \text{Sim}(u, v_i) \mathbb{I}_{\{v \text{ visits keyword } k\}}.$$

It then predicts that u visits locations associated with keyword k if and only if $P(u) \geq \theta$ where $\theta \in [0; 1]$ is a parameter that allows a trade off between accuracy and aggressiveness of the reconstruction technique.

We empirically study the effectiveness of this attack using 1.3 million checkins at 460,663 locations from 40,578 Foursquare users, obtained through crawling publicly available tweets of

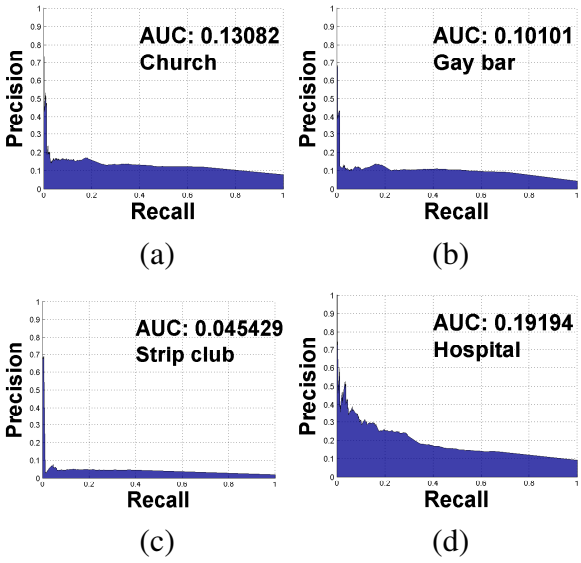


Figure 3: Precision-Recall curves for four sensitive keywords: (a) Church (b) Gay Bar (c) Strip Club (d) Hospital

checkins between March and August 2011. Each Foursquare location is marked with a category, which we assigned to be that location's keyword. In this attack, we consider a severe case where the adversary has compromised 20% of all accounts. We vary the value of θ from 0 to 1 and plot the precision-recall of this attack for various keywords in Fig. 3. As one can see, this attack is rarely effective, even in such extreme case where many user accounts have been compromised. The area under the curve is almost always very small. This turns out to be true even for locations that are sparse, as it is much more difficult to guess right when only a handful of users are visiting a rare location.

This points to an interesting difference between inference in our scheme and de-anonymization attacks. While de-anonymization attacks always benefit from sparsity since the data are present in a sanitized form, in our context, the attack does not always benefit from sparsity. This is because a minimum critical mass of typical behavior is needed in order to run inference. This shows that a proper choice of blacklist could potentially protect many locations, even as several accounts are compromised in the system.

7.3 Attacks on Advertiser Revenue

We now consider if advertisers can unfairly lose money to unscrupulous users of the system. Because users are paid when they are accessed by advertisers, they have an incentive to view or click on many ads, even when they are not interested in the displayed products, to artificially boost their profile's value to derive more money from each click. We label these activities "user fraud."

User fraud is a special case of invalid traffic in online advertising. According to Google's Ad Traffic Quality Resource Center, "invalid traffic includes both clicks and impressions ... [that are] not the result of genuine user interest. This covers intentionally fraudulent traffic as well as

accidental clicks and other mechanically generated traffic."[8] A request for an ad within our system is just like a request for an ad in the current ad ecosystem, but with some privacy-protecting filtering and potential additional location information. Thus, previous techniques used to identify invalid traffic can be used to identify user fraud. There is a lot of recent research on this topic. Dave et al propose methods to fingerprint click spam [12]. Haddadi uses “bluff ads", ads designed to not appeal to humans and thus only be clicked by bots, to defeat click fraud [27]. Information on the structure of Google’s click fraud detection system is available [38]. Beyond academia, multiple startups exist that estimate the rates of click fraud, such as Adometry, Visual IQ, and ClearSaleing (`www.adometry.com`, `www.visualiq.com`, `www.clearsaleing.com`).

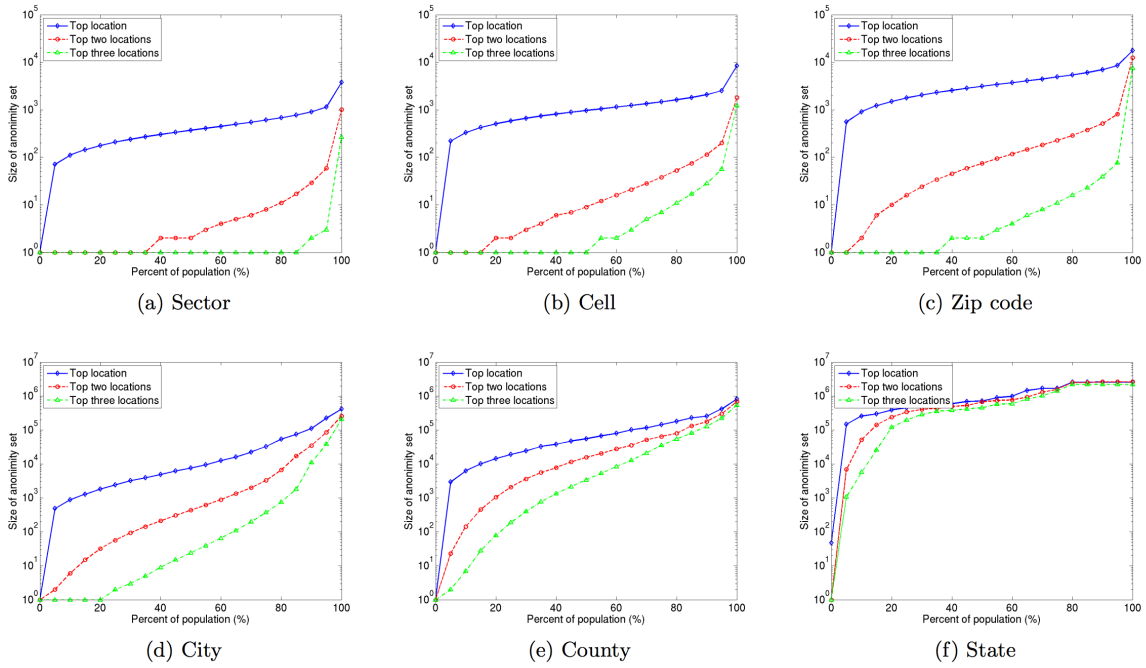
Additionally, it is easier to detect user fraud than traditional invalid traffic because location information is more constrained than web-browsing. Users are physically constrained in how far they can travel in a certain period of time and typically display periodic mobility patterns, returning to their homes at night and spending week days at work locations. A more extreme use of physical constraints would be to use location tags; fingerprints extracted from ambient signals at a specific location at a specific time [54]. These constraints can be used to filter out automated attacks on a system. For example, if a user appears to be traveling faster than is physically possible, we can remove them from the system or verify their accounts with a Captcha or phone call. Because of these physical constraints, and because click fraud prevention techniques can easily be applied to our system, we believe that our system is no more vulnerable to gaming than current online advertising. The ongoing viability of online advertising shows that our solution should likewise not be derailed by invalid traffic.

Beyond automated attacks, users might “physically" attack the system by simply going to a high value location in order to appear more valuable to an advertiser than they actually are. Again, techniques to combat click fraud can be employed here. Click fraud techniques must deal with situations in which users actually click links to unfairly gain money, a nice analogy to this form of attack. Beyond this, traveling to a location takes significant time and effort and will likely be too costly to be a viable way of making money.

8 Related Work

Our work is part of a growing body of work that deals with privacy solutions that aim to reconcile the privacy concerns of users with the economic needs of ‘free’ online web services and mobile applications [24, 25, 62, 72]. Privad [24] and Adnostic [72] are browser based systems that enable behavioral targeting while ensuring users’ PII is not leaked to ad-networks performing the targeting. Our focus in this paper is different – we are concerned with location information on mobile devices. Koi [25] is a system developed to address location privacy by way of location matching – applications and service providers pre-declare which locations they would be interested in and the device releases this information at those specified locations. Our solution is different, in that we have an economic component where application developers need to pay to access the user at the specified location. In addition, neither the device nor applications have to be modified to use our solution. Our work is closely related to transaction privacy [62]. The difference is that we focus on location information for mobile devices and develop a keyword-based disclosure scheme.

[8] `www.google.com/ads/adtrafficquality/index.html`



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

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

Figure 4: Figure from [77] 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.

9 Conclusion

The collection and monetization of location information has become a large concern. The main contribution of this paper is the design and analysis of a solution for location privacy using economics. Our solution is simple – opt-in users decide which locations to reveal and only these locations are sold on an information market. Buyers pay to gain access to users at specified locations. Locations are specified in keywords, a notion intuitive to both end users and advertisers. Our solution relies on a privacy protection component that ensures that the location information the user chooses not to release will not be leaked, and also minimizes the linkage of the user's identity with the released information. Future research directions on keyword-based disclosure may include reducing the role of the trusted third party, larger implementations, and a stronger economic analysis of the solution. A few locations, at a cell level, have been shown to provide poor anonymity [13]. An interesting open question is if keywords provide better k-anonymity.

10 Location Data and Anonymity

10.1 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" [77]. The paper used data from cell phone call detail records (or CDR, see Chapter 2) 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 4. 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.

The Montjoye nature report

10.2 Completed Work

I have investigated the anonymity of location data for users

10.2.1 Linking Users Across Domains with Location Data

Although prior work showed location to be highly *unique* and thus posibly *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. For example, much of cryptography relies on creating highly unique but unpredictable sequences of numbers. To put it more concretely, imagine that each individual had a die with 1000 sides, and each side represented a location. 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.

TODO: put some math here?

Another possibile 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.

TODO: put some math here?

In “Linking Users Across Domains with Location Data", published at WWW in 2016 [63], we tackled this problem, linking users across two entirely different datasets.

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 paramater 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).

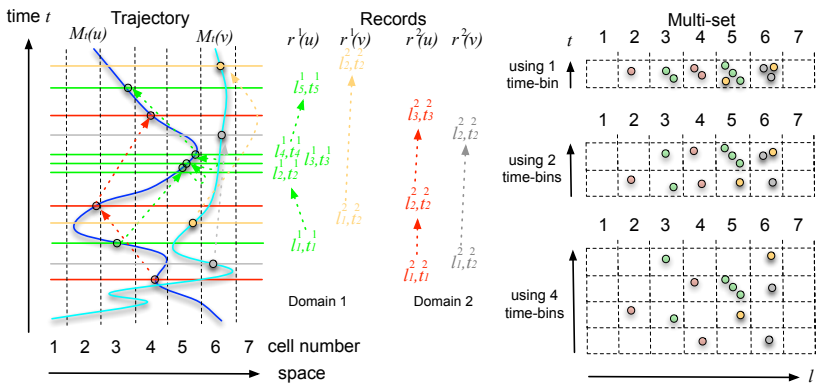


Figure 5: Two space-time trajectories with associated footprints in two domains.

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 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).

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

Figure 5 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 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 [9] and geotagged updates from the microblogging site Twitter [10]. This data was obtained in a prior work by other authors who allowed us to use their data [78]. 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 [11] 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 1.

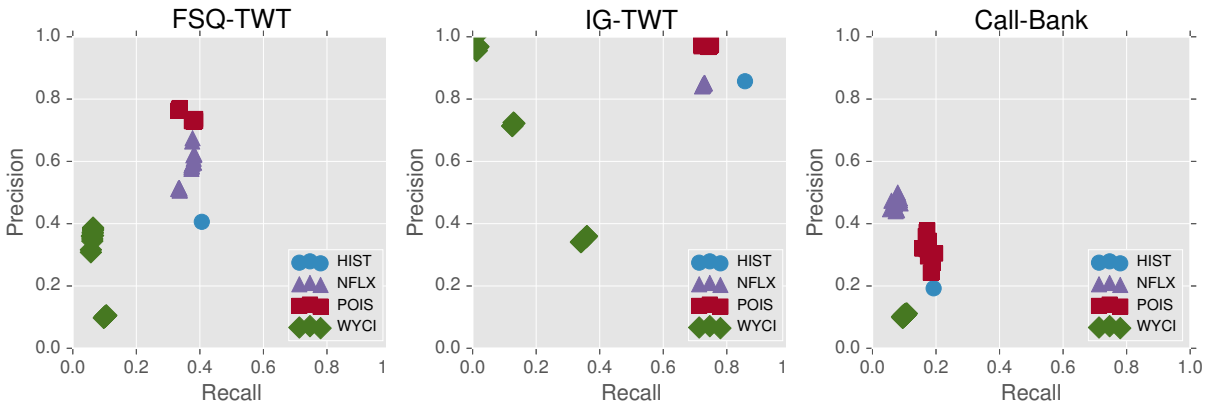


Figure 6: Precision and Recall plots for each dataset.

We now turn our attention to experimental performances of our algorithm. In Figure 6, we show the precision recall plots for our algorithm (for different eccentricity values) and for the other three reconciliation techniques: HIST, NFLX and WYCI. For our algorithm, we used estimated parameters and for the other techniques, we used optimal parameters (found via exhaustive search).

[9] https://foursquare.com/
[10] twitter.com
[11] instagram.com

There are several interesting observations that we can make on Figure 6. 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.

Other results 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.

10.2.2 FindYou: A Personal Location Privacy Auditing Tool

FindYou has two main goals: The first goal of our project is to inform users, regardless of technical skill, about what their location information can reveal. The second goal is to improve research on demographics and mobility by gathering a new dataset with the informed consent of interested users.

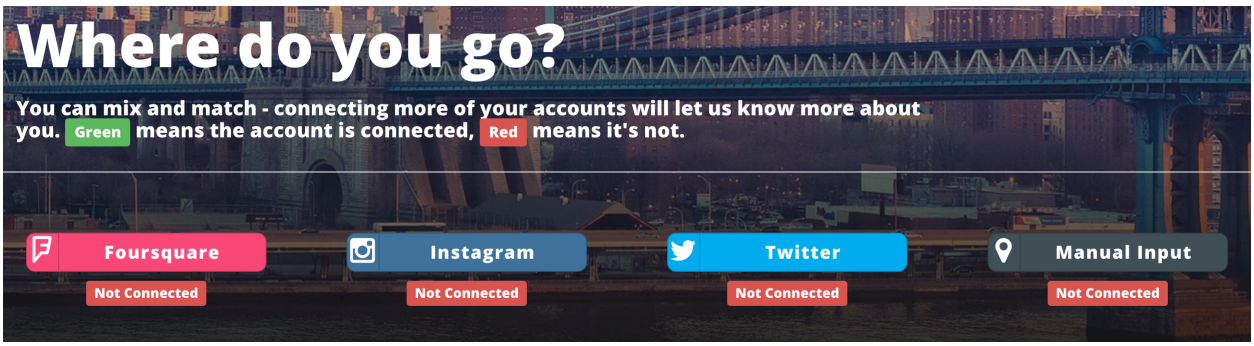


Figure 7: The user is presented with four different ways of connecting his or her location data to the app.

We will begin with a summary of a typical use of 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

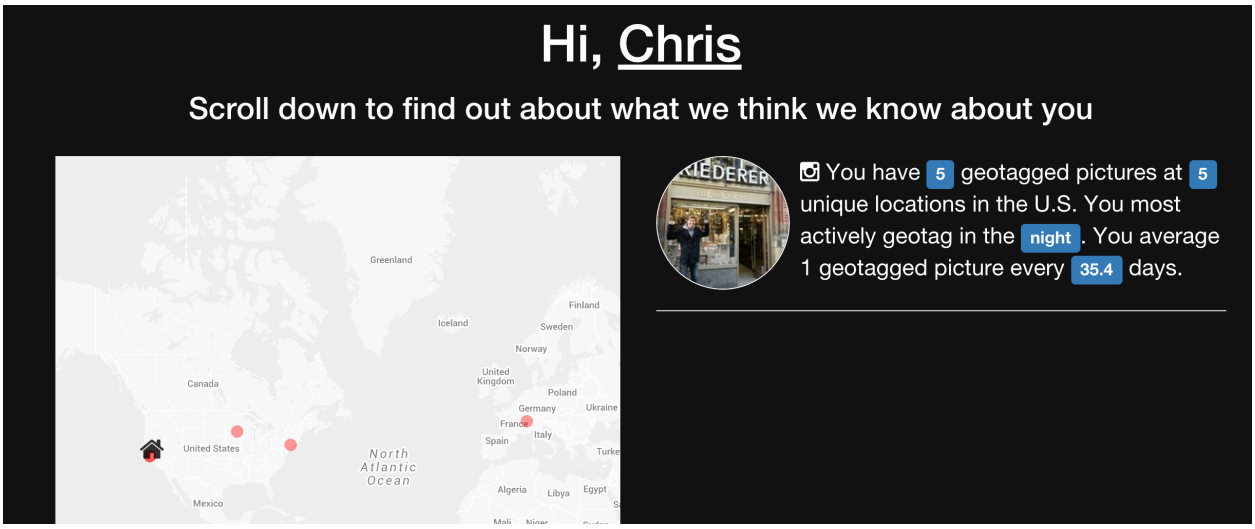


Figure 8: After connecting their data, the user sees an overview of their locations and imported data.

Figure 9: We show a specific guess for the user's home location.

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.

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.

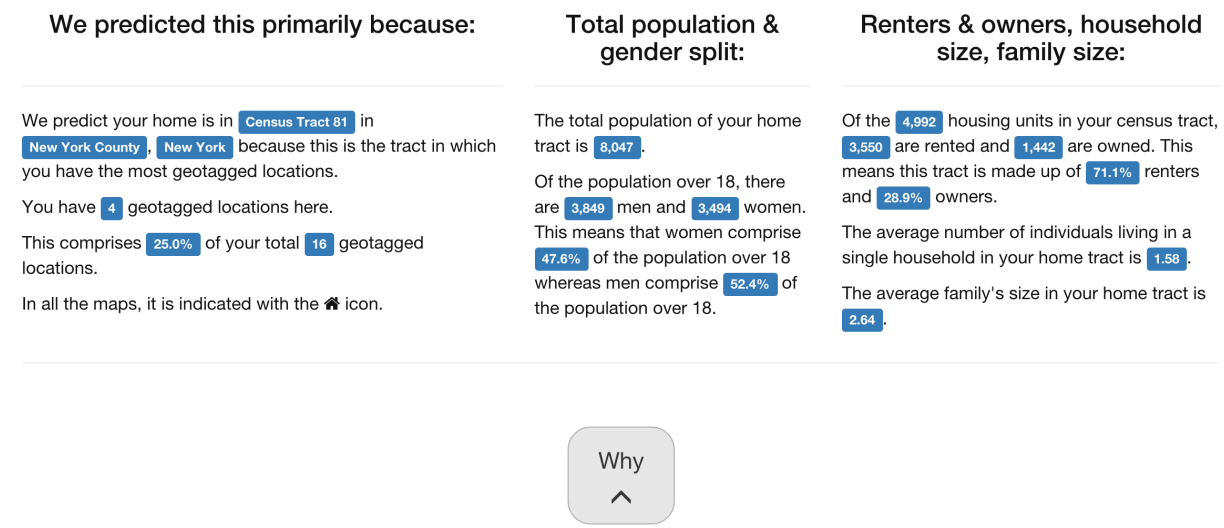


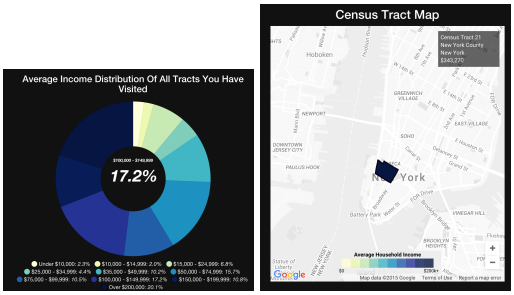
Figure 10: For all predictions, we show additional details about how we made this guess.

Figure 11: The site predicts several demographic attributes, one of which is race. The user has the option to tell us if we are correct.

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.

In the future, we hope to connect more services and also include more advanced location-data



(a) (b)

Figure 12: (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.

uploading. For example, users could include data in standard geographic formats, such as GeoJSON or those used by GIS software. For the time being, we believe that our three chosen services and simple uploading methodology will provide users with an interesting and useful coverage of options.

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.

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

contain 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 bounadires and demographic make-up.

Why only America? Due to our reliance on U.S. Census data, our site currently only bases it's predictions on visits to locations in the United States. We hope to expand to other countries in the future. This presents some challenge, as each census of each country will have different types of data available, different classifications, groupings, and currencies, and different APIs. We look forward to tackling this challenge in future work. For the time being, focusing on the world's third most populous country with one standardized census and many online social network users has appeared to be a good option.

11 Location Data, Demographics, and Bias

11.1 Related Work

11.2 Completed Work

From countries, states, and regions to neighborhoods, local meeting houses, or churches, the places we live and visit are tightly connected to our identities. Location information gathered about an individual over time can reveal much more than a home location. In fact, patterns in human mobility can reveal more private components of an individual's blah, such as their race, gender, or economic status.

In this section, I will discuss two works which explore these issues. Both rely on a dataset connecting images to mobility through geotagged photos on the image-sharing site Instagram. In the first, we label faces for gender, race, and other attributes, and explore whether these features can be inferred from locations visited. In the second, we evaluate the use computational vision techniques to scale up the use of Instagram to connect mobility to demographics for an "automated census".

11.2.1 I DonâĂŹt Have a Photograph But You Can Have My Footprints

Human mobility is intimately intertwined with highly personal behaviors and characteristics. As Justice Sotomayor of the United States Supreme Court stated, "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 [74]." For that reason, previous studies of mobility centered on the risk of either re-identification in sensitive anonymized location datasets or on protecting visits to private locations [13, 26].

However, the re-identification risk based on individual locations is not the only threat. Many users are producing a series of footprints, which might be innocuous individually, however, taken

together can create a sparse yet informative view allowing inferences from their whereabouts. The benefits of revealing locations are obvious: location data can be used for personalizing recommendations [59] and displaying more relevant advertising [45] in order to finance free online services. However, the downsides are more difficult to assess. While an individual data point may create no privacy risk, an aggregated dataset might enable inferences beyond a user's expectation.

In this paper 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. As we discuss in §8, 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. (§7.2).

Our study complements works on human mobility patterns and attribute inference in multiple ways.

First, the use of location data relates our study to previous inquiries into human mobility [10, 22, 55]. In particular, we aggregate location data into mobility patterns and compare our patterns to those published in earlier studies [3, 34, 35] for validation, but furthermore we analyze those patterns both at an individual level and aggregated in multiple demographic groups, including, for the first time, from the perspective of ethnicity. This analysis complements previous studies which have shown that mobility is correlated to social status [9] and community well-being [**?**] measured at city and neighborhood levels. While some studies already demonstrated that mobility traces can uniquely identify individuals [13, 69], the inference of individuals' demographic attributes from location data, that is, the *discriminative* power of location data, remained unexplored. We make inferences beyond trip purpose identification [15], activity type prediction [44, 46], and identification of location types [33].

Previous studies aimed to infer the ethnicities, gender, and other attributes of online users. Often they leveraged linguistic features, such as Facebook or Twitter user names, stated first and last names [8, 52], or Tweet content [59, 60]. Those studies demonstrated an underrepresentation of females and minorities online [52]; a finding which we extend and confirm using photosharing services. Mobility data from mobile phones were used to predict personality traits [14], age [7], and

gender [67], but, in addition to relying on proprietary data, all of these studies solely analyzed call patterns or social network properties as opposed to locations. In contrast, we attempt to infer attributes using *only* location data, making our work more broadly applicable to any technology that can collect mobility information, such as GPS, Wi-Fi, or mobile apps. We additionally examine whether predictions become more accurate with more data, similar to [**?**], and how the granularity of data impacts prediction accuracy.

More generally, our analysis fits into the category of works on extracting information from social networks, such as [11]. Probably, the closest work is [79], which also aims to infer meaning from locations, however, is not concerned with ethnicity. We obtain our data from profiles of the photosharing service Instagram, and our analysis is enhanced with auxiliary information from the geo-social search service Foursquare and the United States Census 2010 [73] (Census). To our knowledge this is the first study demonstrating that it is possible to extract from social networks mobility patterns that are enriched with ethnic or gender information at an individual level. It should be noted in particular that all aforementioned studies of mobile data rely on proprietary data, primarily CDRs, that are only available with the consent of the data owner (e.g., [13, **?**]). In contrast, our methodology is principally reproducible by anyone at a small cost, and our data will be made available shortly after publication. Our study provides a contribution to overcome the lack of publicly available mobility datasets and serves as a validator for their patterns. 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 13.

Data Collection Applying this methodology, 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.

User Labeling To match previous studies [33, 34, 35] that leveraged ZIP codes of CDR billing addresses from the Los Angeles (LA) and New York City (NY) metropolitan areas we randomly chose users from those areas as well. A user's home is the ZIP code where he or she had the most checkins (that is, photos taken). Note that this mitigates the content produced by tourists and other occasional visitors to LA and NY unless those have no other Instagram activity. A combination of workers on Amazon Mechanical Turk (MTurk) and undergraduate students were asked to annotate users' ethnicities and gender based on the users' photos. However, in order to ensure that user pictures on Instagram profiles are sufficient to make a conclusive determination of users' ethnici-

ties and genders we ran a preliminary experiment by selecting 200 profiles at random (excluding celebrities and business accounts) and having each labeled independently by two undergraduate students. We observed a strong agreement on gender (98%). The errors corresponded to a family profile belonging to multiple people and profiles with one picture.

For ethnicity labeling we leveraged Census categories. We asked 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). Merging all remaining Census fields in the last category limits our detail view, although we would otherwise have some annotations being quite rare. 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%. It should be noted that the two labeling students were of different gender and ethnicity themselves. In conclusion, despite sparse data and ethnicity spanning a continuous spectrum, we found that labels are surprisingly predictable and consistent across annotators. As studies confirmed that 91% of teens post a photo of themselves on social networks [48] and that 46.6% of photos are either selfies or show the user posing with other friends [29] there is also evidence in many cases that it is actually the account owner who is shown in the pictures.

To scale our annotation, we asked MTurk annotators to label a larger number of profiles for the same metropolitan areas using the same label categories. For consistency, we did not reuse the profiles used for the preliminary experiment described above. Each profile was labeled by two MTurk annotators. 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 as unsupervised crowd workers on MTurk or similar platforms tend to be less attentive than physically available workers [56], who also have the possibility to ask clarifying questions. We were also careful to not drop any labels to avoid the introduction of a systematic annotation bias. Over two days 117 MTurk annotators participated in our task resulting in 1,015 properly labeled users with the labels shown in Figure 14. On the first day the annotators were compensated $0.10 per annotation and on the second day $0.05. 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 [39]. Generally, values above 0.8 are considered as good agreement, values between 0.67 and 0.8 as fair agreement, and values below 0.67 as dubious [49]. Figure 14 shows that we obtained fair and good agreement and, thus, reliable ground truth for both our ethnicity and gender classifications.

Adding Auxiliary Information We collected auxiliary information from two sources. First, for the comparative analysis of demographic patterns with our data in §11.4 we used data from the Census [73] to associate geographic regions with gender and ethnicity distributions. Throughout the study we use Census-defined geographic granularities, ranging from block groups of 600-3k people to neighborhood tabulation areas (NTAs; 15k people), public use microdata areas (PUMAs; 100k people), and counties with populations of up to 2.6 million. We adjusted the distributions

	Spring		*Winter*	
Statistic	LA	NY	LA	NY
Total Checkins (Total CDRs)	135,503 (74M)	109,506 (62M)	118,446 (247M)	98,286 (161M)
Min. Loc./Day	1	1	1	1
1st Qu. Loc./Day	1	1	1	1
Med. Loc./Day	**1**	**1**	**1**	**1**
(Med. Calls/Day)	**(9)**	**(10)**	**(8)**	**(9)**
(Med. Texts/Day)	-	-	**(4)**	**(3)**
Mean Loc./Day	1.97	2.12	1.96	2.1
3rd Qu. Loc./Day	2	2	2	2
Max Loc./Day	73	62	98	69

Table 2: Statistics of our LA and NY subsets compared to the CDR dataset in [34] (where available, in parentheses). Our calculations do not consider any day where a user had no checkins.

by ethnicity- and gender-specific Internet [19, 47] and Instagram [16] usage numbers. As explained in §11.4 we also took into account that Caucasian Hispanics are often perceived as Caucasian alone [51]. Second, for each checkin we obtained Foursquare information on the ten closest venues. We then used Foursquare's average venue popularities and venue categories as features for our inference algorithms (§7.2) since those features could provide an estimate of the types of places a user would visit. We now present a mobility pattern analysis for various population levels. Our dataset reveals mobility trends similar to those of CDRs (§11.3) and generally represents the adjusted Census population well (§11.4). In many cases we are able to detect differences in mobility patterns between ethnic groups and genders that can be plausibly explained by previous sociological findings (§11.5), and we are also able to detect segregation among ethnic groups (§11.6).

11.3 Mobility Patterns

In order to compare the mobility patterns of our dataset to those in the CDR dataset of [34, 35] we only consider checkins for the years 2011 through 2013 each for the Spring months from March 15 to May 15 and for the Winter months from November 15 to January 31 (the LA and NY Spring and Winter subsets, respectively). Table 2 shows the distribution of the data in our subsets compared to those in the CDR dataset [34]. The mobility traces from our subsets are much more sparse. Most notably, while the CDR dataset has at least eight location points from call activity per day for the median user in LA and NY—and even 12 if text messages are added—the data in all of our subsets account for only one location point for the median user per day.

Another insightful metric for comparing mobility patterns is the *daily range*, defined as the maximum straight line distance a phone has traveled in a single day [35]. Daily ranges are characteristic for mobility because, for example, median daily ranges on weekdays represent a lower bound for a commute between home and work locations [35]. Figure 15 shows a subset of our results. Our ranges are generally smaller than those reported by [34, 35]. However, the general trends in both datasets are similar. Most importantly, people in LA have generally greater ranges than people in NY. Also, in both areas people tend to travel longer during the day than at night.

However, there are also differences: according to our data New Yorkers in the 98th percentiles travel farther than Angelinos.

11.4 Demographic Patterns

As our LA and NY subsets are annotated with ethnicity and gender labels (§**??**) we are able to compare the resulting demographic distributions to the respective Census distributions. However, initial comparisons reveal substantial differences. For example, according to the Census there are more females than males (53% vs. 47%) living in Kings County [73] while our observed label frequencies suggest that there should be substantially fewer (43% vs. 57%). This result is even more surprising as the gender-specific usage rates of Internet (70% vs. 69%) [19] and Instagram (16% vs. 10%) [16] should further increase the percentage of females beyond the Census. However, while 86% of female social network account owners set their profile to private, only 74% of males do so [47]. Adjusting the Census distribution for this difference (as well as for gender-specific Internet and Instagram usage rates) leads to a distribution of females and males (49% vs. 51%) much closer to the distribution we observed for our labels.

Similarly to gender, we make adjustments to the Census distributions for the varying percentages of Internet and Instagram usage rates among different ethnicities as well. However, even then we still observed a substantial Hispanic underrepresentation, which was also observed for the southwest of the United States by [52]. We found this phenomenon difficult to assess, specifically, as ethnicity is not significant for setting a profile private [43], activity levels (posting pictures, etc.) are not lower for Hispanics [71], and our annotation disagreements are not higher when the Hispanic label is involved. However, we believe that the reason for the underrepresentation is the perception of Caucasian Hispanics as Caucasian alone. In a study, six of seven Caucasian Hispanics reported that others see them as Caucasian alone [51]. Therefore, we believe that most Caucasian Hispanics were actually labeled as Caucasian (i.e., our annotators agreed on an incorrect classification). Thus, we adjusted the observed label frequencies by adding to the Hispanic labels a number of labels corresponding to the Census percentage of Caucasian Hispanics and subtracting the same number from the Caucasian labels.

We perform chi square tests for goodness of fit comparing the gender and ethnicity distributions of our labels to the corresponding Census distributions for different levels of granularity. In most cases we obtain a value of $p > 0.05$ and find no evidence to reject the null hypothesis that the observed gender and ethnicity distributions follow the corresponding Census distributions. For example, as shown in Figure 16, for eight out of 11 counties in the NY area our tests resulted in $p > 0.05$ providing no evidence that our multi-category ethnicity distributions deviate significantly from the Census distributions. However, there are also cases with differences. It is no surprise that this is true for the state level as our distributions only cover users from the LA and NY metropolitan areas. However, overall we believe our results suggest that geotag data often replicate demographic trends faithfully.

11.5 Mobility Patterns by Demographic

By combining our methodologies from the previous two subsections we now show the differences in mobility patterns between ethnic groups and between males and females, respectively. In particular, we examine differences in daily ranges, home ranges, and temporal checkin characteristics.

Daily Ranges Figure 17 shows some of our daily range results for ethnic groups and genders based on our sets of labeled users for LA and NY. We obtained the same types of daily ranges as described earlier in Figure 15, however, this time for all days of the year. It is striking that Caucasians generally have a higher maximum daily range than the other ethnic groups. Indeed, a two sample Kolmogorov-Smirnov test reveals that the Caucasian range distribution differs significantly ($p < 0.05$) from the African American and Hispanic distribution. This result illustrates a more general finding: daily ranges of Caucasians often differ significantly from those of minorities. For 44% (8/18) of the comparisons of a Caucasian distribution to a minority distribution (three comparisons for maximum weekday, three for median weekday, three for median at night—each for LA and NY) the difference is significant at the 0.05 level. However, for the comparisons among minority distributions we only find 6% (1/18) to be significantly different from each other.

The differences in ranges by ethnicity can be most prominently observed in the comparisons of Caucasians to African Americans and to Hispanics. However, it should be noted that at night all ethnicities exhibit very similar ranges. This finding stands in contrast to the difference in daily ranges between males and females. In fact, the only statistically significant difference ($p < 0.05$) that we observed between male and female ranges occurs for the median daily ranges at night. As shown in Figure 17, females tend to travel smaller distances at night than males. There are many possible explanations for this phenomenon. One reason could be that women travel fewer times at night due to safety concerns [1] and, consequently, also avoid longer trips. In general, for both males and females—as well as for all ethnicities—we find that our observed daily ranges follow a (skewed) log normal distribution.

Home Ranges In order to evaluate differences in mobility with respect to an individual's home location we complement the analysis of daily ranges with the evaluation of *home ranges*. A home range is a straight line distance between someone's home and another place to which the person travels. Different from daily ranges we calculate the home ranges not on a daily basis, but instead consider all home ranges—whether they were the maximum travel distance for a day or not. Based on a user's home location, as specified in §**??**, we calculate the distance between the home and each checkin for the different ethnic groups and genders. Figure 18 shows the resulting CCDFs for the home ranges of the NY users.

Both graphs show a noticeable decrease around the 2,500 mile mark, which is the distance from NY to major hubs on the West Coast of the United States (most notably LA (2,475 miles), San Francisco (2,563 mi), and Seattle (2,405 miles)). Males and females have very similar home ranges at the edges of the graph. However, females travel farther in the medium home ranges. This finding could be based on the fact that women generally take more often vacations [37] and travel longer distances to work when they are employed full-time [41]. It should be noted that the larger home ranges are not inconsistent with the previous observation of shorter ranges for females at night as that result does obviously not consider ranges during the day. The plot for ethnicity is in line with our previous observation that Caucasians travel farther from home than minorities.

Temporal Checkin Characteristics Beyond spatial differences we explore differences in temporal activity as well. Figure 19 shows histograms for checkins by hour of day. As might be expected, we observe periodic behaviors with low checkin levels between 4–6am and peak levels from 3–8pm. On weekends the lows occur at later times than on weekdays suggesting that users

wake up later on weekends. We also see a dramatic increase in activity after 5pm on weekdays, which could correspond to the time at which many users get off of work. When broken up into Caucasians and minorities, we see fairly similar curves, except with a more pronounced weekday after-work increase for minorities. It could be the case that Caucasians work more often in flexible environments. We observe no substantial differences between genders or NY and LA.

11.6 Ethnic Segregation

Location data are the basis for measuring residential segregation, that is, the degree to which two or more groups live separately from one another in different parts of the urban environment [50]. Trends in residential segregation characterize a group's proximity to community resources (e.g., health clinics) and its exposure to environmental and social hazards (e.g., poor water quality and crimes) [61]. In addition to *residential* segregation we also introduce and evaluate *mobility* segregation, which we understand as the degree to which two or more groups *move* to and from different parts of an area. Mobility segregation allows for a dynamic view of segregation, for example, in order to determine a group's ease of access to community resources away from home.

Methodology Various intersecting dimensions of segregation can be distinguished [50]. We explore two standard measures, each for a different dimension: the interaction index measures the dimension of exposure (the extent to which minority group members are exposed to majority group members in an area [50]) and the entropy index measures the dimension of evenness (the extent to which minority group members are over- or underrepresented in an area [50]). The interaction index, B, can be understood as the probability of a minority group member interacting with a majority group member and is defined [76] by

$$B_{kl} = \sum (\frac{n_{ik}}{N_k})(\frac{n_{il}}{n_i}), \tag{1}$$

where n_{ik} is the population of ethnic minority group k in area i (e.g., in a ZIP code area), N_k is the number of persons in group k in the total population of all areas, n_{il} is the population of ethnic majority group l in area i, and n_i is the area population.

The entropy index was used in social network research before [11] and has the advantage over other indices that it can be used to measure segregation for more than two groups. We define the entropy index [76], H, as

$$H = \frac{H^* - \bar{H}}{H^*}, \tag{2}$$

where H^* is the population-wide entropy defined by

$$H^* = -\sum_{k=1}^{K} P_k ln(P_k), \tag{3}$$

and $\bar{H}$ is the weighted average of the individual areas' entropies defined by

	Hisp./Cauc.		*Af. A./Cauc.*		*Oth./Cauc.*	
Gran.	LA	NY	LA	NY	LA	NY
County	0.29 (-2%)	0.34 (+2%)	0.27 (+1%)	0.3 (-2%)	0.3 (-3%)	0.4 (0%)
PUMA	0.32 (-6%)	**0.39** **(+3%)**	0.43 (+4%)	0.42 (+7%)	0.31 (-10%)	0.49 (+5%)
NTA	- -	0.54 (+6%)	- -	0.43 (+3%)	- -	0.55 (+7%)
ZIP	0.36 (-19%)	0.56 (0%)	0.33 (-23%)	0.55 (+1%)	0.58 (-1%)	0.5 (-7%)
∅ % Diff.	**5%**		**6%**		**5%**	

Table 3: Interaction index (B) for different granularities based on labeled Instagram data. Differences to the interaction index calculated from Census data are shown in percentage points in parenthesis. For example, the probability of a Hispanic person to interact with a Caucasian person on the PUMA granularity level for NY is 39%. However, as shown in parenthesis, this result is an overestimation by three percentage points over the Census distribution probability of 36%. The last row of the table shows the mean difference between our labels and the Census for the three different ethnicities in absolute percentage points for both LA and NY together. Note that NTAs are not available for LA and that we also did not analyze the state level as the label and Census distributions differ significantly (Figure 16).

$$\bar{H} = -\sum_{i=1}^{I} \frac{n_i}{N} \sum_{k=1}^{K} P_{ik} ln(P_{ik}), \tag{4}$$

where K is the number of different ethnic groups, P_k is the proportion of ethnicity k in the total population, I is the number of different areas, n_i is the population in an area, N is the sum of the population from all areas, and P_{ik} is the proportion of the population of ethnicity k in area i (while it is defined that $P_{ik}ln(P_{ik}) = 0$ for $P_{ik} = 0$).

For both interaction and entropy indices we make use of our sets of labeled users for LA and NY, however, exclude all areas for which the label distribution deviated significantly from the Census distribution as indicated by $p \leq 0.05$. Thus, for example, as shown in Figure 16, on the county level we do not include Queens, Kings, and Bergen. These exclusions are necessary as otherwise the accuracy of our results decreases substantially. Recall that we define a user's home as the ZIP code where he or she had the most checkins (§**??**) and that we adjust label and Census distributions (§11.4).

Residential Segregation Tables 3 and 4 show our results for the interaction and entropy indices, respectively. For the most part the interaction between Caucasian and minority group members can be considered fairly high [31]. All three minorities in LA and NY have similar probabilities of interacting with Caucasians. The measurement errors of 5% (Hisp./Cauc. and Oth./Cauc.) and 6% (Af. A./Cauc.) between our labeled data and the Census suggest that our results are overall reliable. The inaccurate results for LA on the ZIP code level appear to have been caused by the smaller number of data points. While the level of interaction seems to increase when areas become

	Entropy				
Metro	County	PUMA	NTA	ZIP	∅ % Diff.
LA	0.01 (-2%)	0.15 (+8%)	- -	0.15 (+9%)	**3%**
NY	0.08 (0%)	0.14 (+1%)	0.08 (0%)	0.09 (+4%)	

Table 4: Entropy index (H) for different granularities based on labeled Instagram data. Differences to the entropy index calculated from Census data are shown in percentage points in parenthesis. As explained in Table 3, the last column shows the measurement error. As further explained in Table 3, we did not consider NTA (LA) and state granularities (LA and NY).

more fine-grained, this phenomenon seems to be caused by the different area coverage for the various granularities. For example, it is not present when considering all NY city areas, where the Census distributions for the interaction of African Americans and Caucasians are: 0.41 (County), 0.25 (PUMA), 0.2 (NTA), and 0.22 (ZIP).

With entropy index scores ranging from 0.01 to 0.15, as shown in Table 4, we find another indicator for low segregation [31]. However, it should be noted that this low level of segregation is a characteristic of the particular areas we investigated. For example, for all NY city areas at the NTA level we calculated an entropy of 0.31 indicating higher segregation. However, with mean differences of 5% (Hisp./Cauc.) and 6% (Af. A./Cauc. and Hisp./Oth.) between the results for our labeled data and the Census-based calculation our findings are generally reliable. As in the case of interaction, we believe that any existing inaccuracies could be due to small numbers of data points.

Mobility Segregation We evaluate mobility segregation based on the same measures as residential segregation—interaction and entropy indices. However, instead of using home locations we leverage checkin data. More specifically, for each user we calculate the percentage that he or she spent at a certain area and sum the resulting values for all users of a certain ethnicity. This method aims to avoid overcounting of active users. Our results are shown in Table 5 and indicate that segregation levels in terms of where people go are similar to levels of where people live. Indeed, it would have been surprising to see higher segregation levels as members of minority groups may work in predominantly Caucasian areas. Furthermore, it would also have been a surprise to see lower levels of segregation as residential segregation is already relatively low.

We now show how location data by itself allows to infer ethnicity and gender of individual Internet users. We introduce a simple frequentist approach (§11.7), describe considerations informing our methodology (§11.8), and present the results of its application (§11.9).

11.7 A Simple Inference Algorithm

Our approach yields two advantages: (1) it provides a formulation of the problem that is intuitive and (2) it remains generic so as to be easily applicable to any sparse location dataset. We use the following assumptions: each user, i, belongs to one of two classes, C_1 or C_2. Class C_1 (respectively C_2) is associated with a probability distribution μ_1 (respectively μ_2) over a discrete set of locations, representing the fraction of time spent by users of that class in that location. Our main assumption is that a user i makes n checkins, denoted $X^{(i)} = (X_1^{(i)}, \ldots, X_n^{(i)})$ at locations that are drawn

	Interaction			*Entropy*
Metro	Hisp./Cauc.	Af. A./Cauc.	Oth./Cauc.	All Eth.
LA	0.55 (+1%)	0.57 (0%)	0.58 (-1%)	0.06 (+1%)
NY	0.54 (-2%)	0.53 (-1%)	0.53 (-5%)	0.06 (+2%)
∅ % Diff.	**1%**	**1%**	**3%**	**1%**

Table 5: Mobility interaction and entropy indices for ZIP code granularity based on labeled Instagram checkin data. Differences to the residential interaction and entropy indices calculated from Census data are shown in percentage points in parenthesis. The last row of the table shows the mean difference between our labels and the Census in absolute percentage points for both LA and NY together.

Task	*Best Algorithm*	*Parameters*	*Important Features*	*Baseline Accuracy*	*Accuracy*	*AUC*	*F1*
Ethnicity NY	Logistic Regression	L1, $C = 0.01$	Avg. ZIP ethnicities	0.52	**0.72**	**0.76**	**0.7**
Ethnicity LA	Logistic Regression	L1, $C = 1$	Avg. ZIP ethnicities	0.50	**0.63**	**0.66**	**0.6**
Gender NY	Logistic Regression	L2, $C = 0.1$	Men's Store	0.53	**0.58**	**0.59**	**0.5**

Table 6: 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.

independently from this user's class probability distribution. The prior probability that a user is in class C_1 or C_2 is denoted π_1 and π_2, respectively.

Note that this model does not use notions of times of the day, geographies, or auxiliary information. It applies to most location datasets as it is agnostic to how they were generated, anonymized, or in which granularity they are available. Such model serves as a starting point to approximate human mobility [23]. However, in practice humans show periodicity [22] or even social bias [10] in their movements, and users in a class may not be identically distributed, which is why it is important to test our technique using real data (§11.9). Under our assumptions, the problem of classifying users in their respective class reduces to a simple hypothesis testing. If i is in class C_1 then for any location l, we have

$$\forall j,\ P(X_j^{(i)} = l | i \in C_1) = \mu^{(1)}(l), \tag{5}$$

so that

$$P(X^{(i)} = (l_1, \ldots, l_n) | i \in C_1) = \mu^{(1)}(l_1) \ldots \mu^{(1)}(l_n), \tag{6}$$

by independence, and applying Bayes' rule

$$P(i \in C_1 | X^{(i)} = (l_1, \ldots, l_n)) = \frac{1}{1 + \frac{\pi_2 \mu^{(2)}(l_1) \ldots \mu^{(2)}(l_n)}{\pi_1 \mu^{(1)}(l_1) \ldots \mu^{(1)}(l_n)}}. \tag{7}$$

The Neyman-Pearson lemma states under the assumptions above that the most powerful statistical test to determine which class a user belongs to from its checkins is the likelihood ratio test. A

maximum likelihood rule classifies a user in class 1 iff

$$\pi_2 \mu^{(2)}(l_1) \dots \mu^{(2)}(l_n) < \pi_1 \mu^{(1)}(l_1) \dots \mu^{(1)}(l_n) \tag{8}$$

or, equivalently, if we have

$$\sum_{k=1}^{n} \ln \frac{\mu^{(1)}(l_k)}{\mu^{(2)}(l_k)} > \ln \frac{\pi_2}{\pi_1} \, . \tag{9}$$

We expect that our predictions are more accurate on users with more checkins. One can show under these assumptions that this classifier's error probability for a user decreases *exponentially* as the number of checkins n grows, that is,

$$P(\text{error}|n \text{ checkins}) \approx_{n \to \infty} 2^{-n\mathcal{C}(\mu_1, \mu_2)}, \tag{10}$$

where μ_1 and μ_2 are the probability distributions associated with C_1 and C_2, and $\mathcal{C}$ denotes the *Chernoff information*, defined as $\mathcal{C}(\mu_1, \mu_2) = -\min_{0 \le \lambda \le 1} \ln \sum_l \mu_1(l)^{1-\lambda} \mu_2(l)^{\lambda}$.

Based on this analysis, a simple algorithm to infer ethnicity or gender can first estimate μ_1, μ_2 and π_1, π_2 using the training data and then classify according to this likelihood rule.

11.8 Methodology

Our purpose is to explore generally what might be inferred about users from their location data only. This affected our methodology in a few key ways. First, we utilized well-understood, commonly-applied techniques that could easily be employed by anyone with access to mobility data. We also used publicly available data-sources. Second, to make our results applicable to other sources of location data beyond Instagram, we did not use features specific to Instagram, such as the social network graph or user-generated descriptions. Thus, our work should be viewed as a lower-bound on the accuracy of what can be inferred using location data. Adversaries with access to more detailed auxiliary information, more data about each user (such as a contact list or recent purchases), or more advanced machine learning techniques might achieve better results.

We considered two questions: (1) Can minorities be distinguished from Caucasians? (2) Can women be distinguished from men? We represented users as feature vectors, using three classes 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" (the collection of which we explained in §**??**); 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 [57] and tested the algorithms logistic regression, decision trees, naive Bayes, and support vector machines (SVMs). As a baseline, we predicted ethnicity or gender based on the class distribution, giving us baseline accuracies of 52% for ethnicity in NY, 50% for ethnicity in LA, and 53% for gender in NY.

Auxiliary Data Auxiliary information about a location derived from Foursquare or the Census may not always be available, e.g., in countries without publicly available census data or when locations are anonymized. Furthermore, a labeled training set of user data may not always be available either. To understand the performance of an algorithm that does not have access to any

data beyond counts of visits to locations, we applied our **Bayesian** algorithm to our data. To test if labeled data was necessary to guess ethnicity, we developed a simple decision rule that used no labels. Based on Census data we calculated the average percentage of Caucasians living in all locations that a user visited. If this percentage was over the metropolitan area's average, we predicted that the user was Caucasian. If it was below, we predicted that the user was of a minority ethnicity. We called this the **Unsupervised Threshold** algorithm. We compared this algorithm to an algorithm with access to labeled data, which learned an optimal threshold rather than using one derived from publicly available Census data and which we dubbed the **Supervised Threshold** algorithm. Finally, we compared these algorithms against our best performing algorithm, run with all features at the lowest granularity. We call this the **Full** algorithm.

Data Granularity The granularity of location data can vary greatly depending on how it is created. Previous research has investigated the impact of location granularity on anonymity [13, 77]. To investigate the impact of granularity on inferences, we represented our location data at several different granularities defined by the Census ranging from block groups to states. The ethnic makeup of a large granularity area, such as a county, will typically be more similar to the overall metropolitan area's ethnic makeup than a small granularity area like a city block. Thus, increasing the granularity should make inferences more difficult.

Data Quantity Finally, with four different analyses, we studied the impact of data quantity on prediction accuracy. First, to explore the impact of user activity on inference accuracy, we grouped users according to their number of geolocated Instagram photos. Next, we investigated the impact of location diversity by grouping users according to the number of distinct ZIP codes they visited. 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 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.

11.9 Results

The results of our best-performing algorithms are displayed in Table 6, and a detailed comparison of accuracy as a function of granularity can be seen in Figure 20. Our results suggest that geotag data can be used to infer an individual's ethnicity and gender. The accuracy for predicting ethnicity falls squarely within what has been reported for other types of datasets. On the lower bound, in their work of predicting individual Twitter users as African American or not based on linguistic features of Tweets [59] report as best performance an F-1 score of 0.66. On the upper bound, for predicting whether the ethnic origin of a phone user is inside or outside the United States based on a rich feature set containing Internet usage, call, text message, and location features [**?**] achieved an F-measure of 0.81 and for gender an F-measure of 0.61. For gender [79] achieved an F-measure of 0.81 for social network users in Beijing and 0.82 for Shanghai based on spatial, temporal, and location context knowledge. Given that our dataset contains far fewer features our results demonstrate that geotags are surprisingly powerful in predicting gender and ethnicity.

Auxiliary Data It can be observed in Figure 20 that the Supervised Threshold algorithm performs much better than the Unsupervised Threshold algorithm suggesting that labeled data improves the algorithmic accuracy across the board by roughly 5%. Interestingly, the Bayesian algorithm performs comparably to the Supervised Threshold algorithm. Thus, an algorithm with no semantic information about visited locations performs just as well as one that knows the ethnic makeup of all visited locations. This suggests that an adversary with enough location data labeled with demographic data could obtain reasonable levels of accuracy with no knowledge of what locations were visited. Even if locations are "anonymized," that is, GPS coordinates or venue names were obscured, they can still be used to infer demographic information about the user.

Data Granularity The Full algorithm (that is, our best performing algorithm, with access to all features at all levels of granularity) achieves the best performance; no algorithm with access to restricted, coarser-grained features is as accurate.

The performance of all algorithms decreases at the most coarse granularities. This is most likely because the ethnicity distributions of larger regions are closer to the overall distribution of the metropolitan area and provide less information. Several algorithms improve in performance at medium granularities, such as ZIP and neighborhood. This is most likely caused by the sparsity of our dataset at the most detailed granularity as many blocks are only visited by a few users.

Data Quantity It appears that the accuracy of ethnicity prediction improves with the total number of checkins a user has made as shown in Figure 21. The distinct number of ZIP checkins of a user provides a separate measure of user activity as a user could have a large fraction of checkins in few ZIP codes. We can observe a substantial boost in accuracy after a user checked in at 12 distinct ZIP codes.

We also found that when a user is only observed in a limited set of locations, the inference accuracy increases fast with a relatively small increase in the number of locations. Moreover, it is not even required to focus on the most significant locations of a user to get good inference accuracy. Observations of a user in a few random locations at the tract or neighborhood level might be enough for predicting ethnicity, and those locations may be even selected randomly and must not be necessarily related to the user's most significant places. These results, which are displayed in Figure 22, suggest that inference for the purpose of ethnicity identification is quite robust to data sparseness and obfuscation methods.

This study highlights the risks and opportunities of discriminative big data analysis by demonstrating that it is possible to infer Internet users' ethnicities and genders based on location data *alone*. It also shows that mobility patterns can be studied using publicly available data. Internet users may often be unaware that releasing such data could also disclose possibly sensitive personal information. Simply reducing granularity proved to be insufficient to prevent such privacy leakage as mobility remains discriminative. However, the trove of geotagged pictures available through individual online profiles also yields important insights for beneficial uses, for example, by city planners and social scientists.

As our dataset is similar, both demographically and mobility-wise, to other datasets as shown in §**??**, we believe that our results are generalizable and applicable to other unlabeled datasets. Although it could be claimed that our data is biased by the fact that the users in our study have willingly disclosed their gender and ethnicity by publicly using Instagram, we want to stress that

it would be difficult and possibly unethical to create a labeled dataset of users who *do not* want to disclose their gender and ethnicity.

This work motivates multiple avenues of further research: First, it enables the extension of demographic mobility analysis to many researchers using shareable public datasets and reproducible results. Beyond ethnicity and gender, attributes such as age, occupation, and other lifestyle features may be extracted from users' pictures, and naturally there are many other mobility properties to account for beyond, for example, daily ranges. Second, better understanding the discriminative power of location data might inform the design of tools for raising user awareness about the information they reveal. This insight motivates revisiting mobility modeling and the inferences it renders possible to empower users to make at will their locations as clear as a photograph or as opaque as footprints in the mud.

11.9.1 Scaling up the Census with Social Media

The growth of publicly available online information from social networks provides many opportunities to demographic researchers. However, this data is often messy and unlabeled, causing researchers to need to label profiles with demographics. This labelling is either done manually, a costly and time-consuming process, or done via automated algorithms. The inputs to these algorithms have previously been text-based, taking a user's posts or name as input. Techniques involving computational vision have been left unused, due to a lack of training data from users and problems of low accuracy.

In this work, we evaluate the feasibility of combining modern facial recognition techniques with publicly available social media images 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. We conclude with ideas for future improvements and research.

Introduction

The great wealth of publicly available, online social networking data has been a boon to demographic research due to its richness and scale. Never before has such an amount of human behavioral data been easily obtained and analyzed. However, before this data can be used to study demographics, each user must be labeled with demographic categorizations. This poses a particular challenge in many online social networking (OSN) sites which often do not display or even obtain the demographic information of its users.

To meet this need, researchers in the past have tried a variety of techniques. Manual labeling by individually investigating each profile is costly in terms of time, effort, and money. Some studies have relied on data provided by marketing companies or data aggregators [21, 4]. Due to cost and issues of reproducibility, these sources of data are not available to all researchers.

To improve the scale of labeling while keeping costs low, researchers have used automated techniques which range in complexity. For example, researchers have compared public names to lists of gender and ethnicity for those names [52, 8]. Others have run simple algorithms on location data [64] or more sophisticated techniques that incorporate text posted and the structure of a user's social network [59, 58]. These techniques offer some promise but often are not very robust and may not be applicable to OSNs with little textual interaction, such as Instagram.

Researchers that use these automated tools for labeling must be wary of introducing algorithmic

bias. As argued in [68], data mining can lead to biased results. Algorithms that have disparate accuracies in demographic labeling could cause erroneous or biased results.

In the past, computational vision has not been an effective technique for labeling the demographics of social media users, due to three issues: (1) CV algorithms being too slow, (2) CV algorithms having low accuracy, (3) a lack of publicly available and uniformly popular photographs as data. However, in recent years, face recognition tools have improved to become both highly accurate and efficient. Additionally, the social network Instagram has made image-sharing a ubiquitous activity across most of the developed and much of the developing world.

Instagram is an interesting social network to study for a variety of reasons. With over 400 million users at the time of writing, over 5% of the world's population user Instagram. A quarter of these users are based inside the United States, meaning that nearly 1 in 3 United States citizens uses Instagram [32]. Beyond its scale, Instagram is interesting for its content. Photographs are extremely rich, capturing information on all sorts of human activities and interactions. Although images can be more difficult to analyze than text, the research community has begun to study Instagram behavior [30, 2] and even selifes [70].

In this paper, we show that a popular face recognition API can be used to scalably and accurately learn the gender and race of Instagram users. We use a dataset of 200 Instagram profiles, labeled for gender and race, to analyze the practicality and accuracy of facial recognition, achieveing 86% accuracy for gender and 82% accuracy for race in a limited setting. We additionally explore the accuracy of labeling different demographics as the amount of data increases, showing some concerns about algorithmic bias. We conclude with ideas for future work.

Data

Collection

We used a subset of the Instagram data collected in [64]. In this paper, the authors 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. They then collected the user IDs of users who had commented or liked his photos, gathered their metadata, and repeated this process. A subset of these users were then selected based on geography– only users with more than half of their photographs taken in Los Angeles or New York were kept.

Two research assistants labeled a randomly selected subset of 200 of these profiles for gender and race. After filtering for private, deleted, or business profiles, 172 profiles remained. For gender, the labelers selected from male, female, or other. In practice, only the male or female categories ended up being used. For race, a subset of the United States Census categories were used: White, Black, Hispanic, Asian, and other. The labelers agreed on gender for 170 of the profiles and for race on 147. The process resulted in 76 profiles as Male and 94 as Female. For race, 75 were labeled White, 28 as Hispanic, 27 as Black, 16 as Asian, and 1 as other.

Our next step was to recognize and label faces present in these Instagram users' profiles using computer vision. For this task, we used Face++ [18], a popular API with reported high degress of accuracy [2]. In addition to recognizing faces within images, Face++ labels race from White, Black, Asian with a *confidence score* 0-100, gender from Male, Female with a *confidence score* 0-100.

For each user in this data set, we gathered the metadata of the first 100 Instagram photos. Each image was then analyzed with the Face++ API. Face++ only requires that a URL to an image is passed to it. Therefore, this methodology does *not* require that any images are downloaded, uploaded, or even viewed by a human labeler. Note that not every photo on Instagram has a face

in it, and some have more than one.

This resulted in 170 distinct users with at least one or more face present in their photos. We obtained 5,272 photos and depicting 12,143 faces. Additionally, for each user, we passed the URLs for their profile pictures. A total of 70 users had profile pictures in which Face++ could detect a face. 73 faces were found: 2 profiles pictures had 2 faces in them.

Description

We next compared several Instagram behaviors by demographic. Figure 23a shows that women in our dataset typically have more faces in their photographs than men. There appears to be more parity in number of faces for each of the considered racial groups, with Hispanics and Asians displaying slightly more faces in their photos.

We examine potential differences in the total number of photos posted by a user to their account in Figure 24. There does not appear to be a great difference for any particular group.

We investigate the relationship between gender, race, and following behavior in Figure 25 . Instagram has a directional follower relationship, akin to that of Twitter. If a user, Alice, follows Bob, it means that Alice will see updates from Bob in her feed. Bob, however, will not see updates from Alice unless he decides to follow her. Figures 25a and 25c show differences in following behavior within our dataset for gender. It appears that men have more followers, and follow slightly more people.

In Figures 25b and 25d, we see can make a few observations for following behavior among race in our dataset. First, we see that the upper 50% of Black users in our dataset have more followers than other groups, except for at the very top percentiles, where White users dominate. At the highest percentiles, Black users *follow* the most people.

Methodology The face recognition software only works at the level of a single photograph. We thus need to use an algorithm to go from the data of each picture to labeling an entire profile. We rely on one main assumption: the owner of profile will appear in more photos than any other individual.

This assumption led us to test several different algorithms:

- **Majority rule**: Count each face as a vote. Profile is labeled with the gender (or race) with the most votes.
- **Weighted majority rule**: Count each face as a vote. However, a face now gets as many votes as the confidence score. Thus, faces with lower confidence get lower weight. Profile is labeled with the gender (or race) with the most votes.
- **Profile picture**: We simply take the result of the profile picture labeling.

Additionally, we applied a “face weight" correction to all of these. In a photo with three faces, only one of these faces could be the user. It therefore might make sense to weight each face in this photo lower than in a photo with one face. The face weight (fw) correction does just this, multiplying the weight of photo by the inverse of the number of faces in that photograph. In our original example, the weight of each of the three faces would be multiplied by one third. This is equivalent to averaging the gender or race (potentially with confidence scores) of all users in a photograph.

Results

Gender Our human labelers categorized 76 profiles as Male and 94 as Female. Running our three algorithms with and without the face weight correction, we obtained the following results:

- **Majority rule**: 85.1%
- **Weighted majority rule**: 85.7%
- **Majority Rule, FW**: 86.9%
- **Weighted Majority Rule, FW**: 87.5%
- **Profile picture**: 86.3%

We observe that using the face weight correction improves the results for both algorithms. Although profile pictures provide accurate information, only 70 out of 170 of these users had a profile picture that included a face. For the remainder of this section, we will focus on the highest accuracy algorithm, weighted majority rule with FW. A feature of this algorithm is that it outputs a probability for each user, enabling us to analyze the performance in more detail.

In Fig. 26, we group individuals on how what our algorithm predicts is the likelihood that that individual is female, and plot the accuracy within each group. This shows us how well calibrated our algorithm is; if the output probability estimates are perfectly accurate, then half of the users with estimate 50% should be female, and this plot should align on the 1:1 line. Instead, we see that the line is above the diagonal, meaning our accuracy estimate is actually an underestimate. Most likely, we could incorporate a prior probability into our algorithm to make it better calibrated.

In Fig. 27, we plot an ROC curve, showing the trade-off between accurate and inaccurate labelings when using a threshold on the algorithm's output probability. For example, if we label as female all users with a probability of female over 60%, around 75% of female users would be correctly labeled, and we would exclude properly all but around 7% of males.

In Fig. 29, we observe that for users labeled female, as more faces are detected in their profiles, accuracy increases. Perhaps counterintuitively, we see a dip for men, where male users with the 40-80% most faces have much lower accuracy those in the bottom two quintiles, the bottom 40%. One possible hypothesis to explain this is that as some users add more faces, they will start to add a larger diversity of faces. When this diversity increases, accuracy may decrease. We leave the question of whether this is the mechanism open for later work.

Race

For race, 75 users were labeled White, 28 as Hispanic, 27 as Black, 16 as Asian, and 1 as other. However, Face++ only labels users as Asian, Black, or White and will therefore always be incorrect on any of our users categorized as Hispanic or "Other". Thus, we'll present results both for all users, and for the reduced set of users labeled manually by our research assistants as Asian, Black, or White ("filtered"). Running our algorithms with and without the face weight correction, we obtained the following results:

- **Majority rule**: 64.4% (all) 79.3% (filtered)
- **Weighted majority rule**: 66.2% (all) 82.1% (filtered)
- **Majority Rule, FW**: 66.4% 82.2% (filtered)
- **Weighted Majority Rule, FW**: 66.2% (all), 82.1% (filtered)

- **Profile picture**: 57.1% (all), 71.1% (filtered)

On the users for which we have some hope of accuracy, the best algorithm achieves 82.2% accuracy. For the remainder of this section, we will constrain our results to the Weighted Majority algorithm, FW, due to the probabilities that it outputs and its nearly identital performance to the next best algorithm. Additionally, we will look at some labelings as a binary classification problem between White users and Minority users.

Based on Fig. **??**, the algorithm does not appear to be well-calibrated, both in being overconfident in the users with low probability estimations, and being underconfident in users with higher estimates.

An important aspect of demographic labeling is considering issues of the digital divide or disparate impact. In Fig. 30, we see that accuracy is much lower on minorities than it is on White users. Again, we see a lowering of accuracy as the number of faces increases, akin to the dip in accuracy Male users show in Fig. 29.

Discussion

Among our dataset, we see some examples of differences in behavior. For example, women tended to have more faces in their photos, and black users tended to have more followers. A larger sample and careful statistical analysis should be taken to verify these results. Applying machine learning to many more profiles could reveal, and possibly help explain, these and other differences in behaviors on Instagram.

In applying these techniques there are dangers of algorithmic bias. The large difference in accuracy between white users and minority users is an example of this. Our technique as it stands could suffer from this issue. For example, users with more diverse faces in their profile (both gender and racially) may be harder to label. Using a thresholding technique to only obtian high-accuracy users might leave only users who display strong homophily.

Further exascerbating the problem is that racial minority users actually *decrease* in accuracy as there is more data about them, up to a certain point. Clearly more work is need to be done to understand this issue. It may be important to consider the trade-offs here. A more accurate algorithm may not be preferable to an algorithm that is equally accurate across demographic groups, and that behaves similarly in regards to the scaling of data.

Conclusion

In this work, we've shown that computational vision techniques have some promise in becoming a valuable tool for demographers. By combining facial recognition with the OSN Instagram, we've proven that the race and gender of users can be inferred with high precision and recall.

We see several important next steps to this work. First, a larger scale verification of the results of this work should be obtained, with more users, aiming for diversity in many senses– culturally, economically, racially, geographically, etc. Such a verification should investigate the accuracy of the technique on various demographics in order to minimize algorithmic bias.

Another direction is to use more powerful machine learning techniques on this problem. For example, instead of naively incorporating all faces in a profile, a researcher could cluster faces based on a similarity score. The largest cluster would most likely contain the users face. Other facial recognition software packages, with a wider range of races, or with other features, could improve upon these results.

Finally, this technique could be used to engage in studies of various demographic groups and answer different questions. Do different demographics use social networks in different ways?

What can we learn about interactiokn on the OSN between groups? Combining demographic data with location data, we could additionally learn about immigration or human mobility.

12 Proposal Topic I

As described in Chapter 11, 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 differnet 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. I will examine the trade off between recommendation accuracy, bias, and revenue.

Over the course of several months I have gathered photo metadata from the popular image-sharing application Instagram. I have run these photos through a program that recognizes faces within each image, tagging it with age, gender, and ethnicity. This will create the largest publicly available dataset that I know of connecting human mobility to demographics.

Machine learning systems utilize location in making recommendations. However, location can be highly correlated with potentially sensitive traits, such as ethnicity. I plan to look at

The project will emerge in several stages.

1. Collection of instagram data (completed).
2. Labeling of instagram data with Face++ API (completed).
3. Initial analysis and descriptive statistics of dataset (in progress).
4. Full problem specificiation: algorithms, inputs, and objectives.
5. Apply de-biasing to algorithms and analyze impacts.
6. Create recommendations for algorithm designers.

13 Proposal Topic II

In Chapter **??**, I proposed analyzing debiasing algorithms in the setting of an typical online for-profit company trying to optimize their profit. Beyond private enterprise, algorithms play an important role in the civil domain, from decisions about whether to release prisoners on bail to the hopefully fair allocation of scarce resources. The purpose of this project is to take an in-depth look at a government-run matching algorithm, the New York City High School Assignment, with an aim towards analyzing and possibly mitigating inequality.

The New York City Department of Education has a large challenge in efficiently and fairly placing TODO(a large number of) students into high schools. The Department uses a matching algorithm which has some successes: 92% of students are matched and 85% are assigned to one of their top five choices. At the same time, New York City schools are highly racially segregated, with around half of all schools having a student body that is over 90% black and Latino, despite the city's overall student population being just TODO% black and Latino. There are a variety of potential explanations for this result. For example, are the rank lists of students self-selecting into racially homogeneous schools? New York housing has high levels of de facto segregation, and thus students only ranking and attending schools near their homes could be another cause. Additionally, decision criteria at schools, a lack of opportunities at lower levels, or other factors could be causes.

The purpose of this research is to understand if different populations of students are exhibiting different behaviors in their rank lists, and to what extend these differences lead to the skewed results we see in practice. I intend to analyze several different groups, such as racial groups and economic groups. Beyond analyzing the match data, I will additionally adapt and apply Dwork's fairness algorithm **??** and analyze the impact on student utility, school utility, and segregation.

To conduct this research, I will need data from the New York City Department of Education, namely:

1. The rank lists and assignments of students who entered the High School Admissions Program, as well as the rank lists for the schools.

2. Biographic dataset files for the anonymous students, which includes information on age, ethnicity, free lunch status (an indicator of socioeconomic status), attendance data, and more.

3. If available, normalized information about the admissions criteria or requirements associated with each school.

There are two main deliverables for this work: data analysis for hypothesis testing, and an analysis of a debiasing algorithm. In the hypothesis testing portion, I will examine if there are differences in rank-list creation across racial groups and socioeconomic groups. The biographic dataset, available from the DoE for researchers, contains information about ethnicity, language spoken at home, and a commonly-used proxy for socioeconomic status: student entrance in reduced or free lunch program. Student are only eligible for reduced or free lunch if they live in a household with annual income below a certain threshold (TODO(verify, get numbers)). I will look for differences in the following behaviors:

- Length of rank list
- Average school quality of rank list
- *Distribution* of school quality on rank list
- Geographic distribution of schools
- Current racial/socioeconomic make up of school

In the analysis of the debiasing algorithm, I will first adapt Dwork's fairness algorithm **??** to work with matching data. Dwork's algorithm relies on the existence of a similarity metric between

users. I will develop a metric (based on standardized criteria as test scores, attendance, etc.). There are many possibly metrics TODO(mention some) and I plan to test out several. Dwork's paper additionally provides a method of "fair" affirmative action. A measure of utility for students can be calculated as matched school quality or matched school rank on rank list. A school's utility can be calculated as the school's average ranking of its matched students. I plan to test several similarity metrics and affirmative action techniques and investigate the impact of utility for both schools and students.

This work is entirely contingent upon the availability of this data. As such, this project proposal is given as a possible additional undertaking, and will not form the core of my thesis given the high level of risk. A number of other researchers at Columbia have previously obtained this data from the NYC Department of Education. I submitted a formal data request to the Department of Education on February 9th, 2017, and hope to hear back soon.

14 Research plan

References

[1] Emily Badger. This is how women feel about walking alone at night in their own neighborhoods. http://www.washingtonpost.com/blogs/wonkblog/wp/2014/05/28/this-is-how-women-feel-about-walking-alone-at-night-in-their-own-neighborhoods/, May 2014.

[2] Saeideh Bakhshi, David A Shamma, and Eric Gilbert. Faces engage us: Photos with faces attract more likes and comments on instagram. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 965–974. ACM, 2014.

[3] Richard Becker, Ramón Cáceres, Karrie Hanson, Sibren Isaacman, Ji Meng Loh, Margaret Martonosi, James Rowland, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. Human mobility characterization from cellular network data. *Communications of the ACM*, 56(1), January 2013.

[4] Bin Bi, Milad Shokouhi, Michal Kosinski, and Thore Graepel. Inferring the demographics of search users. In *WWW '13*, 2013.

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

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

[7] Jorge Brea, Javier Burroni, Martin Minnoni, and Carlos Sarraute. Harnessing Mobile Phone Social Network Topology to Infer Users Demographic Attributes. In *SNAKDD'14: Proceedings of the 8th Workshop on Social Network Mining and Analysis*. ACM Request Permissions, August 2014.

[8] Jonathan Chang, Itamar Rosenn, Lars Backstrom, and Cameron Marlow. epluribus: Ethnicity on social networks, 2010.

[9] Zhiyuan Cheng, James Caverlee, Kyumin Lee, and Daniel Sui. Exploring millions of footprints in location sharing services, 2011.

[10] Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *KDD '11: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM Request Permissions, August 2011.

[11] Justin Cranshaw, Eran Toch, Jason Hong, Aniket Kittur, and Norman Sadeh. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, UbiComp '10, pages 119–128, New York, NY, USA, 2010. ACM.

[12] V. Dave et al. Measuring and fingerprinting click-spam in ad networks. In *ACM SIGCOMM*, 2012.

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

[14] Yves-Alexandre de Montjoye, Jordi Quoidbach, Florent Robic, and Alex (Sandy) Pentland. Predicting personality using novel mobile phone-based metrics. In *Proceedings of the 6th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, SBP'13, pages 48–55, Berlin, Heidelberg, 2013. Springer-Verlag.

[15] Z. Deng and M. Ji. *Deriving Rules for Trip Purpose Identification from GPS Travel Survey Data and Land Use Data: A Machine Learning Approach*, chapter 72, pages 768–777. 2010.

[16] Maeve Duggan and Joanna Brenner. The demographics of social media users - 2012. *Pew Research Center*, 2013.

[17] William Enck et al. Taintdroid: an information-flow tracking system for realtime privacy monitoring on smartphones. In *USENIX OSDI*, 2010.

[18] Face++. The face++ api. In *www.faceplusplus.com*, 2016.

[19] Thom File. Computer and internet use in the united states. http://www.census.gov/prod/2013pubs/p20-569.pdf, May 2013.

[20] Al Franken. Location privacy protection act, 2011. www.govtrack.us/congress/bills/112/s1223.

[21] S Goel, JM Hofman, and M. Irmak Sirer. Who Does What on the Web: A Large-scale Study of Browsing Behavior. *Proceedings of the International Conference Weblogs and Social Media (ICWSM)*, 2012.

[22] M González, C Hidalgo, and Albert-László Barabasi. Understanding individual human mobility patterns. *Nature*, 2008.

[23] M Grossglauser and D Tse. Mobility increases the capacity of ad hoc wireless networks. *Networking, IEEE/ACM Transactions on*, 10(4):477–486, 2002.

[24] Saikat Guha et al. Privad: practical privacy in online advertising. In *USENIX NSDI*, 2011.

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

[26] 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.

[27] H. Haddadi. Fighting online click-fraud using bluff ads. In *ACM CCR*, pages 22–25, 2010.

[28] P Hornyack, S Han, J Jung, S Schechter, and D Wetherall. These aren't the droids you're looking for: retrofitting android to protect data from imperious applications. *Proceedings of the 18th ACM conference on Computer and communications security*, pages 639–652, 2011.

[29] Yuheng Hu, Lydia Manikonda, and Subbarao Kambhampati. What we instagram: A first analysis of instagram photo content and user types, 2014.

[30] Yuheng Hu, Lydia Manikonda, Subbarao Kambhampati, et al. What we instagram: A first analysis of instagram photo content and user types. 2014.

[31] John Iceland, Daniel Weinberg, and Lauren Hughes. The residential segregation of detailed Hispanic and Asian groups in the United States: 1980-2010. *Demographic Research*, 3:593–624, 2014.

[32] Instagram. Instagram press page. In *https://www.instagram.com/press/?hl=en*, 2016.

[33] 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.

[34] 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.

[35] 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.

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

[37] Kelton. 4th annual springhill suites annual travel survey. http://news.marriott.com/springhill-suites-annual-travel-survey.html, April 2013.

[38] Carmelo Kintana et al. The goals and challenges of click fraud penetration testing systems. In *ISSRE*, 2009.

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

[40] Balachander Krishnamurthy, Delfina Malandrino, and Craig E. Wills. Measuring privacy loss and the impact of privacy protection in web browsing. In *Proc. SOUPS*, pages 52–63, New York, New York, USA, 2007. ACM Press.

[41] Mei-Po Kwan. Gender, the home-work link, and space-time patterns of nonemployment activities. *Economic Geography*, 75(4):pp–370, 1999.

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

[43] Kevin Lewis, Jason Kaufman, and Nicholas Christakis. The taste for privacy: An analysis of college student privacy settings in an online social network. *J. Computer-Mediated Communication*, 14(1):79–100, 2008.

[44] Lin Liao, Dieter Fox, and Henry Kautz. Extracting places and activities from GPS traces using hierarchical conditional random fields. *Int. J. Rob. Res.*, 26(1):119–134, January 2007.

[45] Jack Lindamood, Raymond Heatherly, Murat Kantarcioglu, and Bhavani Thuraisingham. Inferring private information using social network data. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 1145–1146, New York, NY, USA, 2009. ACM.

[46] Feng Liu, Davy Janssens, Geert Wets, and Mario Cools. Annotating mobile phone location data with activity purposes using machine learning algorithms. *Expert Syst. Appl.*, 40(8):3299–3311, June 2013.

[47] Mary Madden. Privacy management on social media sites. *Pew Research Center*, 2012.

[48] Mary Madden, Amanda Lenhart, Sandra Cortesi, Urs Grasser, Maeve Duggan, Aaron Smith, and Meredith Beaton. Teens, social media, and privacy. *Pew Research Center*, 2013.

[49] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[50] Douglas S. Massey and Nancy A. Denton. The dimensions of residential segregation. *Social Forces*, 67(2):281–315, 1988.

[51] Sara McDonough and David L. Brunsma. Navigating the color complex: How multiracial individuals narrate the elements of appearance and dynamics of color in twenty-first-century america. In Ronald E. Hall, editor, *The Melanin Millennium*. Springer, Dordrecht, 2013.

[52] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist. Understanding the Demographics of Twitter Users. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11)*, Barcelona, Spain, July 2011.

[53] 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.

[54] Arvind Narayanan et al. Location privacy via private proximity testing. In *NDSS*, 2011.

[55] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. An empirical study of geographic user activity patterns in foursquare, 2011.

[56] Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision Making*, 5(5):411–419, 2010.

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

[58] Marco Pennacchiotti and Ana-Maria Popescu. Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 430–438. ACM, 2011.

[59] Marco Pennacchiotti and Ana-Maria Popescu. A machine learning approach to twitter user classification, 2011.

[60] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents*, SMUC '10, pages 37–44, New York, NY, USA, 2010. ACM.

[61] S. F. Reardon. *A Conceptual Framework for Measuring Segregation and its Association with Population Outcomes*, chapter 7, pages 169–192. John Wiley Sons, San Francisco, CA, USA, 2006.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] John T. Roscoe and Jackson A. Byars. An Investigation of the Restraints with Respect to Sample Size Commonly Imposed on the Use of the Chi-Square Statistic. *Journal of the American Statistical Association*, 66(336):755–759, December 1971.

[67] C Sarraute, P Blanc, and J Burroni. A study of age and gender seen through mobile phone usage patterns in Mexico. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 836–843, 2014.

[68] Andrew Selbst and Solon Barocas. Big Data's Disparate Impact. *PLSC Conference Communication*, pages 1–57, May 2014.

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

[70] Flávio Souza, Diego de Las Casas, Vinícius Flores, SunBum Youn, Meeyoung Cha, Daniele Quercia, and Virgílio Almeida. Dawn of the selfie era: The whos, wheres, and hows of selfies on instagram. In *Proceedings of the 2015 ACM on conference on online social networks*, pages 221–231. ACM, 2015.

[71] Statista. Social networking time per user in the united states in july 2012, by ethnicity (in hours and minutes). http://www.statista.com/statistics/248158/social-networking-time-per-us-user-by-ethnicity/, 2012.

[72] V Toubiana, A Narayanan, and D Boneh. Adnostic: Privacy preserving targeted advertising. *Proc. NDSS*, 2010.

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

[74] 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)).

[75] Wall Street Journal. Apple, google collect user data, 2011. `http://on.wsj.com/gDfmEV`.

[76] Michael J. White. Segregation and diversity measures in population distribution. *Population Index*, 52(2):198–221, 1986.

[77] 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.

[78] 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.

[79] Yuan Zhong, Nicholas Jing Yuan, Wen Zhong, Fuzheng Zhang, and Xing Xie. You are where you go: Inferring demographic attributes from location check-ins. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 295–304, New York, NY, USA, 2015. ACM.

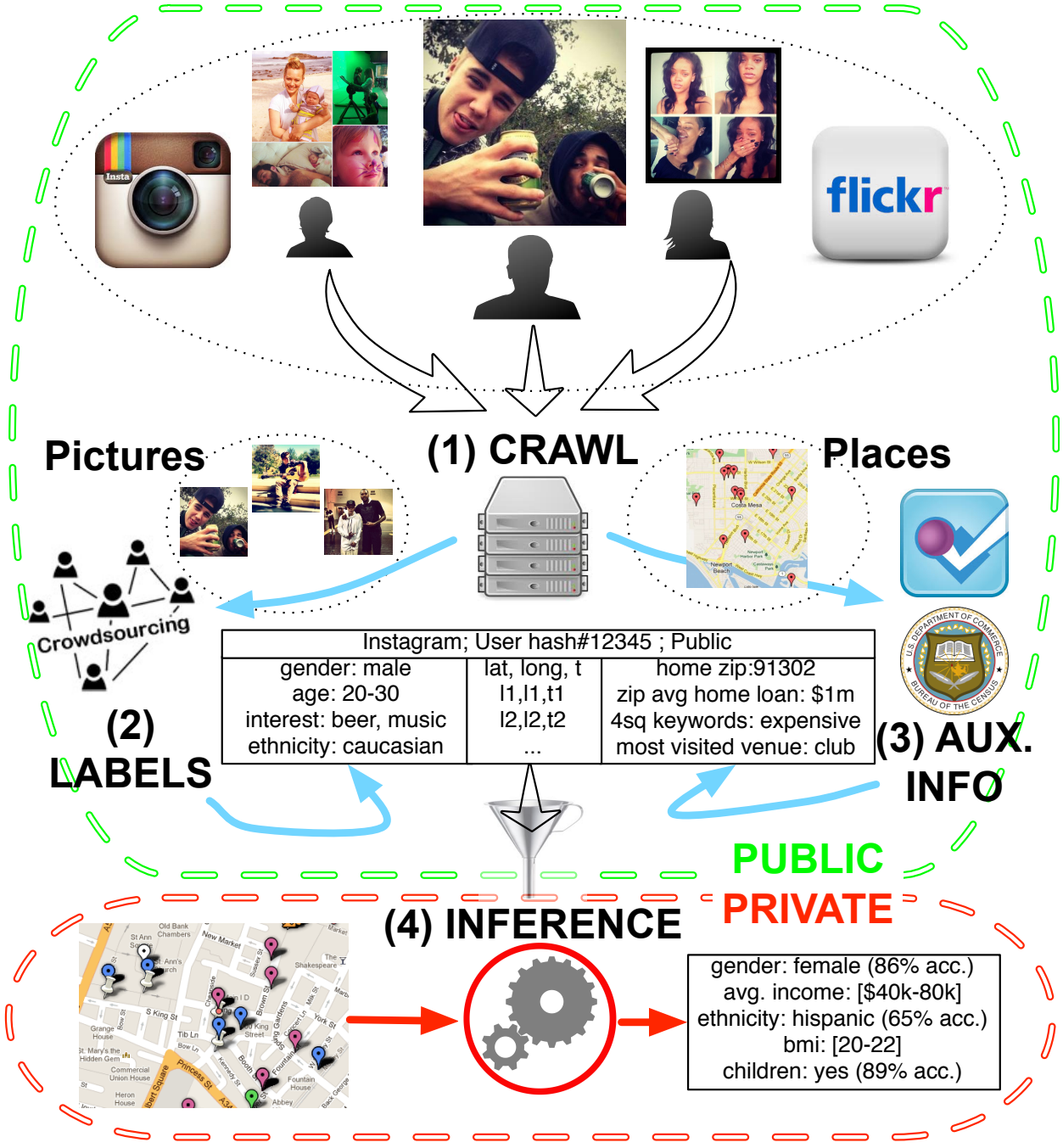
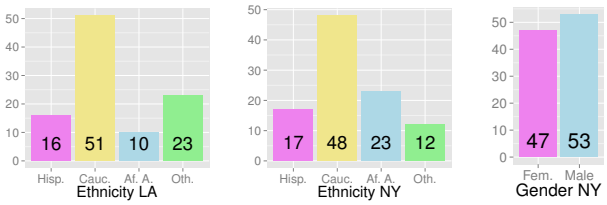
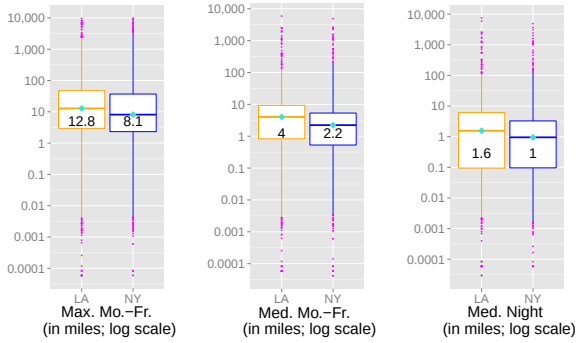


Figure 13: Methodology overview. A mobility dataset can be built in the following steps: (1) Public user profiles of a photosharing service are crawled and photo metadata are extracted into a database (Data Collection). (2) Corresponding photos are labeled (with labels for ethnicity, gender, etc.) by crowd workers in an online labor marketplace (User Labeling). (3) The dataset is further enhanced with auxiliary data, e.g., with the information that a certain location is close to a restaurant (Adding Auxiliary Information). (4) The dataset can then be used to analyze attributes on various demographic levels or train and test classifiers for individual inferences.



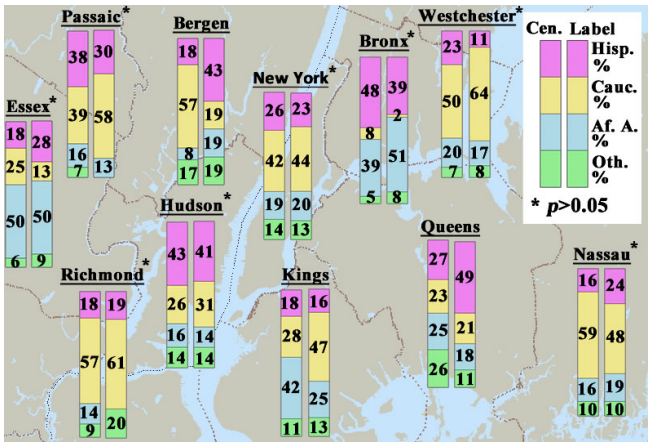
	Ethnicity LA	*Ethnicity NY*	*Gender NY*
Users (n)	427	588	241
K.'s α Multi.	**0.74**	**0.68**	-
K.'s α Binary	**0.78**	**0.74**	**0.85**

Figure 14: Annotations for LA and NY. Top: percentages of user labels for the different categories. Bottom: absolute numbers of labeled users and annotation agreement results.



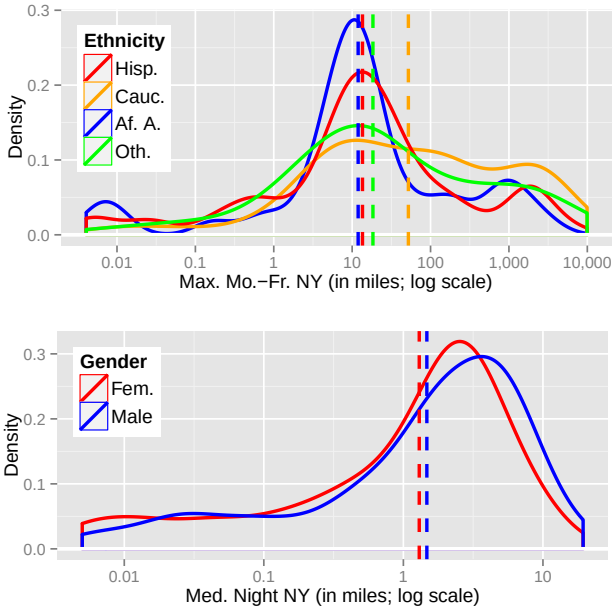
	Max. Mo.–Fr.		*Med. Mo.–Fr.*		*Med. Night*	
%	LA	NY	LA	NY	LA	NY
98	2,471.7 (2,467)	3,625.6 (2,455)	133 (32)	209.9 (29)	117.4 (23.1)	129.9 (19.4)
75	47.5 (130)	37 (111)	9.3 (10)	5.3 (8.2)	6.1 (8)	3.3 (5.6)
50	**12.8** (36)	**8.1** (27)	**4** (5)	**2.2** (3.8)	**1.6** (4)	**1** (2.6)
25	3 (17)	2.3 (12)	0.8 (2)	0.5 (1.3)	0.1 (1.4)	0.1 (0.7)
02	ϵ (1.6)	ϵ (1.3)	ϵ (0)	ϵ (0)	ϵ (0)	ϵ (0)

Figure 15: Daily ranges in miles. Top: boxes show the 25th, 50th, and 75th percentiles; whiskers the 2nd and 98th percentiles. Bottom: table with the percentiles represented in the boxplots. The maximum range (Max. Mo.–Fr.) is a user's longest distance and the median range (Med. Mo.–Fr.) a user's median distance, each taken on a single day for the entire Spring subset on a weekday [35]. The median range at night (Med. Night) represents the median distance a user has traveled on a day for the entire combined Spring and Fall subset from 7pm–7am [34]. Previous results [34, 35] are shown in parentheses. Our calculations do not consider any day where a user had a zero range, that is, had multiple checkins at the same location or a single checkin only. We define $\epsilon < 0.005$ miles.



	Ethnicity Multi-Cat.		*Ethnicity Binary*		*Gender*
Gran.	LA	NY	LA	NY	NY
State	0/1 (0%)	0/1 (0%)	1/1 (100%)	0/1 (0%)	1/1 (100%)
County	1/2 (50%)	**8/11 (73%)**	2/2 (100%)	6/8 (75%)	4/4 (100%)
PUMA	12/16 (75%)	11/17 (65%)	2/2 (100%)	5/6 (83%)	1/1 (100%)
NTA	- -	9/16 (56%)	- -	7/7 (100%)	2/2 (100%)
ZIP	3/3 (100%)	8/14 (57%)	1/1 (100%)	3/3 (100%)	- -

Figure 16: Chi square goodness of fit test results for ethnicity and gender at various levels of Census-defined granularity. Top: detailed view of the multi-category ethnicity distributions for the NY county level. Left bars show the Census distributions (Cen.) and right bars the label distributions (Label). Bottom: complete results of the chi square tests. NTAs are specific to NY and not available for LA. Below the ZIP code and NTA levels we did not have enough data to perform chi square tests. We follow [66] and require the average expected frequency for a chi square test with more than one degree of freedom to be at least two and for a test with one degree of freedom to be at least 7.5. To prevent skewing due to small sample sizes we also use a Monte Carlo simulation with 2,000 replicates.



	Max. Mo.-Fr. NY				*Med. Night NY*	
%	Hisp.	Cauc.	Af. A.	Oth.	Fem.	Male
98	2,480.8	6,509.4	2,270.9	6,788.1	9.8	11.5
75	50.8	592.3	44	187	3.2	4.7
50	**13.5**	**52.1**	**11.9**	**18.4**	**1.8**	**1.9**
25	4.9	7	5.5	3.7	0.4	0.6
02	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ

Figure 17: Daily ranges in miles. Top: density plot of the maximum daily ranges by ethnicity. Middle: density plot of the median daily ranges at night by gender. Bottom: table with the percentiles of the daily ranges represented in the plots. We rounded extremely small daily ranges up to 0.005 miles. Our calculations do not consider any day where a user had a zero range, that is, had multiple checkins at the same location or a single checkin only. We define $\epsilon < 0.005$ miles.

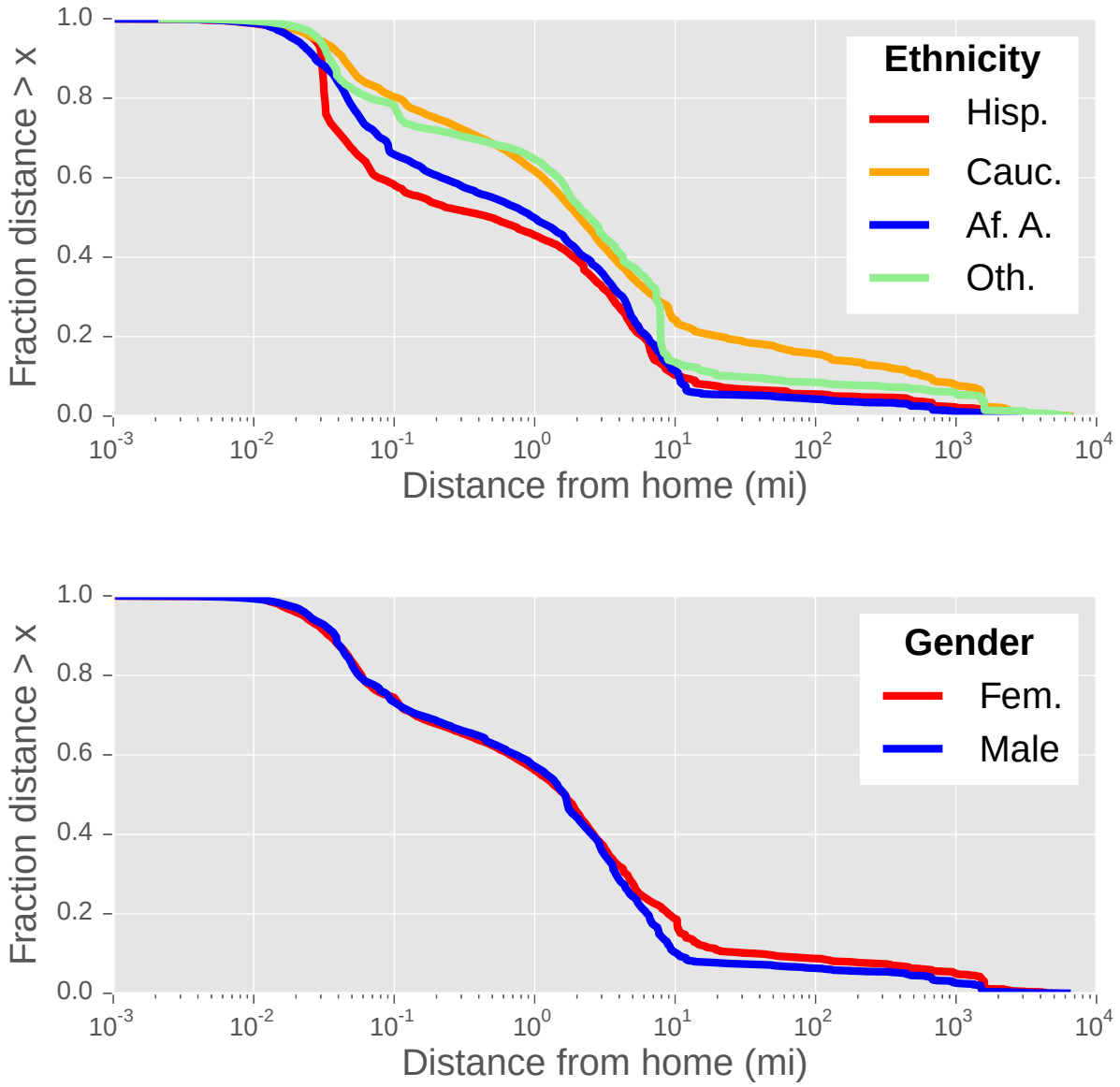


Figure 18: CCDFs of home ranges for NY. Top: CCDFs for different ethnic groups. Bottom: CCDFs for males and females.

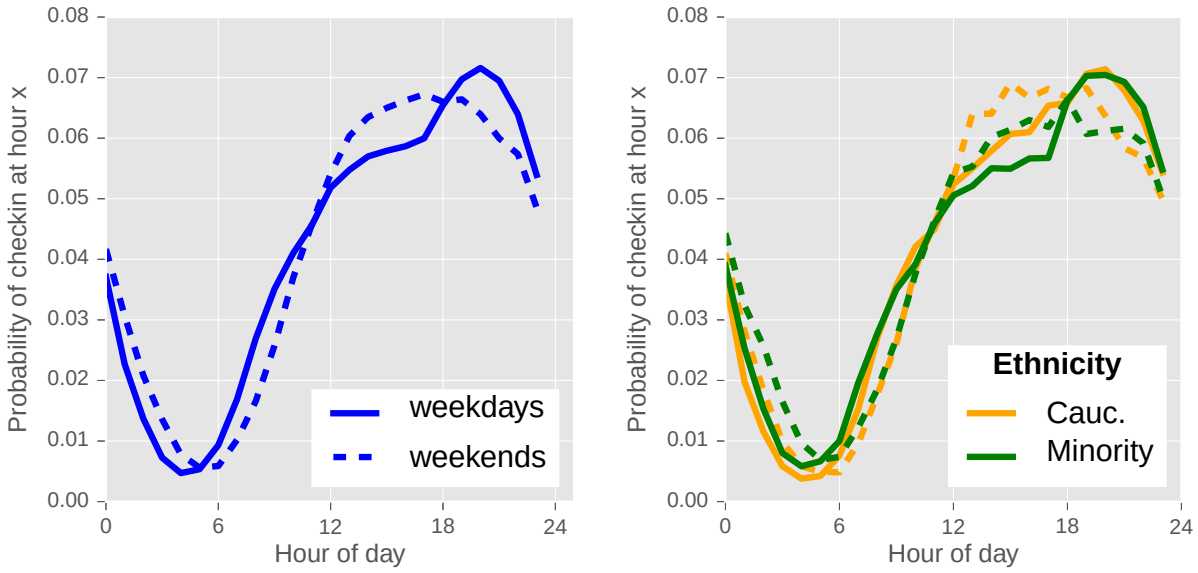


Figure 19: Histograms of checkin times for NY. Left: Comparison of weekends and weekdays for all user groups. Right: Comparison of Caucasian and minority user groups for weekends and weekdays. Dashed lines correspond to weekends, solid lines to weekdays.

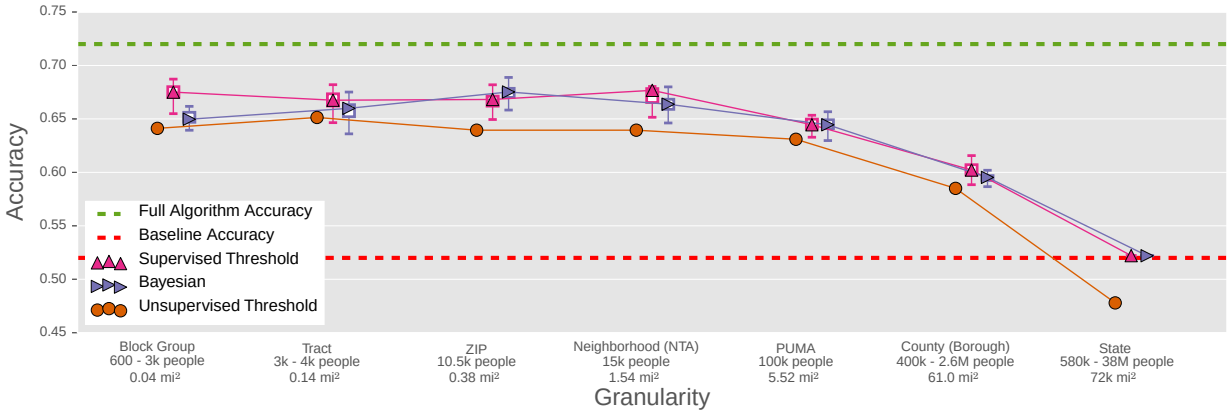


Figure 20: Accuracy of ethnicity prediction versus granularity for our NY population using several different inference techniques. Accuracy increases slightly at the ZIP code and neighborhood granularities and then decreases. Interestingly, the Bayesian algorithm, which uses only counts of visits to locations, performs comparably to the Supervised Threshold algorithm, which uses data on the ethnicity of visited locations.

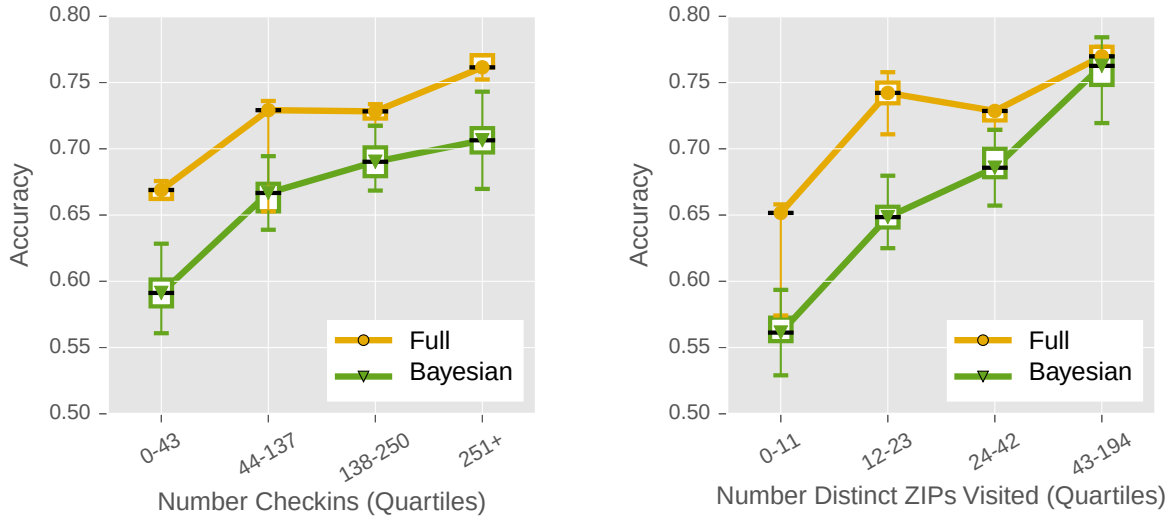


Figure 21: Checkin user activity. Left: accuracy as a function of total number of checkins at ZIP code locations. Right: accuracy as a function of number of checkins at distinct ZIP code locations.

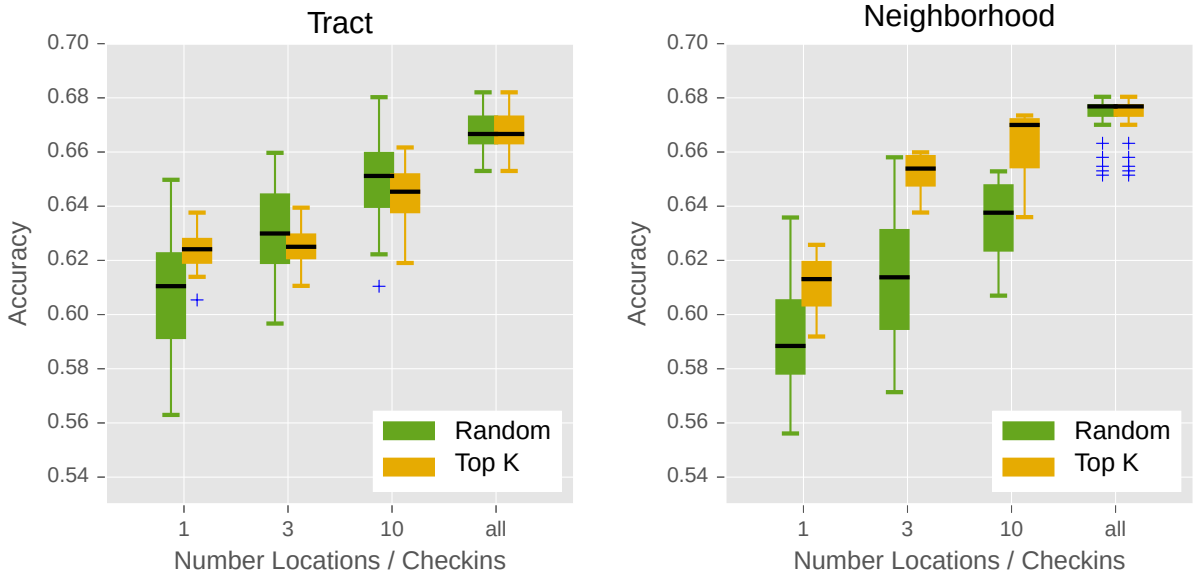
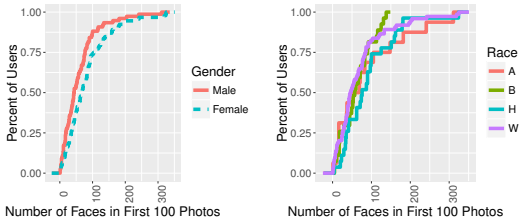
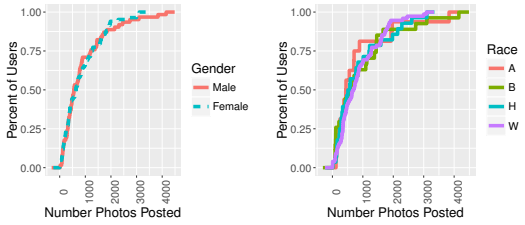


Figure 22: Accuracy of predicting a user's ethnicity from a small number of locations chosen either as most frequently visited locations or randomly. The algorithm used is the Supervised Threshold algorithm. Left: tract granularity. Right: neighborhood granularity.



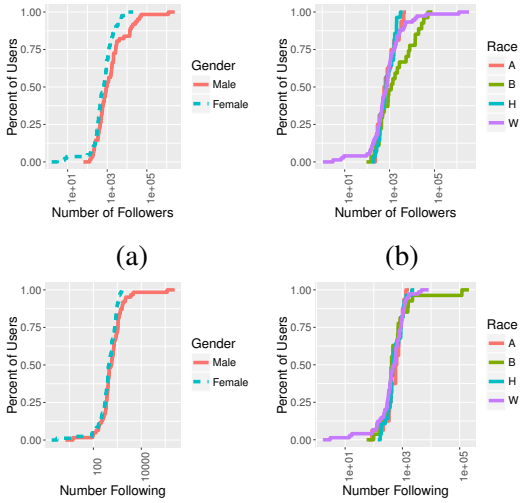
(a) (b)

Figure 23



(a) (b)

Figure 24



(a) (b)

(c) (d)

Figure 25

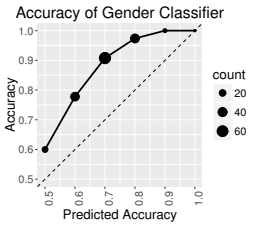


Figure 26: Predicted accuracy versus actual accuracy

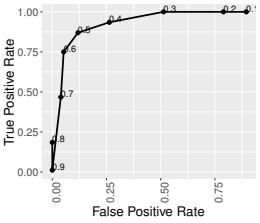


Figure 27: True positive versus false positive rate when labeling at various threshold levels.

Figure 28

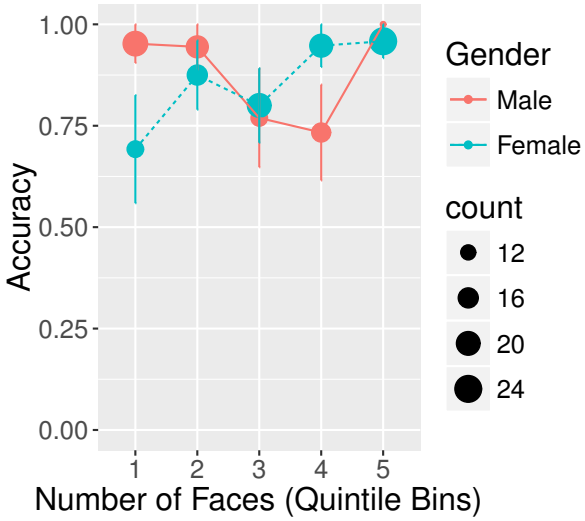
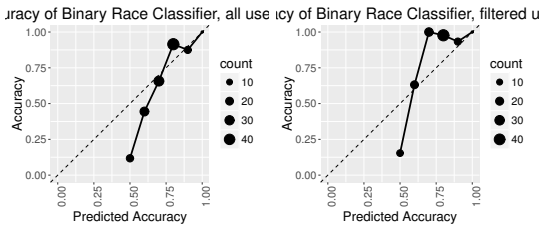


Figure 29: Number of faces detected in first 100 photos vs. accuracy, by gender



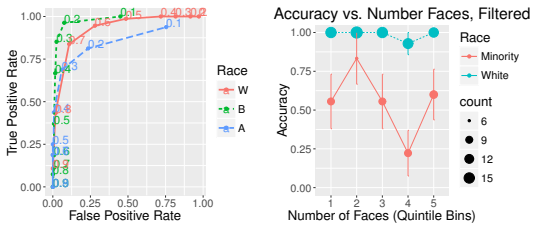

True Positive Rate
False Positive Rate
Race
W
B
A
Accuracy vs. Number Faces, Filtered
Accuracy
Number of Faces (Quintile Bins)
Race
Minority
White
count
6
9
12
15

Figure 30