# Location Data: Perils, Profits, Promise

Christopher Riederer

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2020

ABSTRACT

# Location Data: Perils, Profits, Promise

Christopher Riederer

Most of the modern online economy is based on websites offering free services and content in exchange for advertising access and user data. Web companies collect vast troves of data about their users in order to better target their advertisements. An important subset of this harvested data is the locations visited by users. Location data is valuable as it is a “real world” signal compared to online behaviors: a visit to a store is a stronger signal than a visit to a website, and location data can reveal user attributes that are interesting to advertisers.

The collection of this data, however, raises many concerns. Location data can reveal important attributes that users may not wish to disclose: ZIP codes can reveal income and race, visits to places of worship may allow discrimination, and insurers may want to know about trips to hospitals. The risks exist at both an individual level, with location tied to physical safety, and at a collective level, with inference about group membership a necessary step towards discrimination.

In this thesis, I examine issues of privacy and fairness in the use of location data. In the first portion, I empirically demonstrate new attacks on the anonymity and privacy of users, including a theoretical basis for user identification. In the second portion, I propose and analyze new solutions for dealing with privacy, anonymity, and fairness in the collection and use of location data. In contrast to previous work which presents privacy in abstract ways or ignores the power of data aggregators, the work presented here focuses on concretely informing users and incorporates the economic incentives driving privacy and fairness concerns.

# *Contents*

# *List of Tables*

# *List of Figures*

# *Acknowledgements*

If it takes a village to raise a child, it takes a thriving metropolis to train a Ph.D. I had the fortune of a metropolis populated with some very smart and very compassionate individuals.

This work would not have been possible without my family: my parents Marilyn and Stephen Riederer and my sister Julie. My family has given me many gifts, from an education to vacations to some killer pinochle skills. Probably the biggest gift, which I don't always do a great job of accepting, is the belief that it is important to both work hard and be kind.

Throughout (almost) all of my Ph.D., Michelle Liu has been my rock. She never fails to make me smile, cheer me up, or even laugh at my jokes. She's opened me to new worlds, inspired me with her quiet perseverance, and given me gentle pushes when I needed it. Thank you, Michelle!

My adviser Augustin Chaintreau has shown me that it's possible to do great work while being a great person. In addition to his technical guidance over the years, showing me the craft of research, he has always prioritized humans. I'm wowed by his ability to see the big picture while also supporting us as people, not just workers. Augustin has put up with my desire to do many internships, to leave for a semester to work in politics, and has encouraged me to travel. I believe Augustin took a risk with me as a student. I hope that risk has paid out well!

I've been fortunate to collaborate with and work as an intern for many outstanding researchers and people. Jake Hofman and Dan Goldstein were wonderful mentors to

me at Microsoft Research, New York. I'm indebted to them for introducing me to the wonderful world of dplyr and the tidyverse and in being great teachers to me, being inspirations for me to be a better teacher and communicator. Vijay Erramilli at Telefonica made the first internship of my research career a highly successful one (not to mention fun). Balachander Krishnamurthy at AT&T has provided great advice and been fun to work with. Silvio Lattanzi and Nitish Korula at Google were great collaborators. I had a blast interning at Google with Alex Fabrikant, James Cook, and Boris Smus. I thank Sharad Agarwal and Jitendra Padhye of Microsoft Research Redmond for my time there.

Along the way I've also worked with some great students. Yunsung Kim wowed me with both his abilities and work ethic. Many of my papers have relied upon the help of Danny Echikson, Stephanie Huang, Coralie Phanord, and Jake Cahan.

My fellow Computer Science Ph.D. students at Columbia have collaborated with me, helped me come up with new ideas, lent sympathetic ears, and shown me how to fix the air conditioning in my office. In particular, thanks to Arthi Ramachandran, Mathias Lécuyer, Sebastian Zimmeck, Karl Stratos, Avner May, Jon Bell, Ana-Andreea Stoica, Swapneel Sheth, Etienne Vouga, and many others. Outside of Columbia, I have many friends who kept me sane over the last six years. In particular, thanks to Dan Baneman, Reuben Fischer-Baum, Sarah Brady Stack, Dave and Dan Robinson, Callie Lowenstein, Jon Zytnik, John Acheson, Yuri Shadunsky, Arpit Gupta, Dana Levin-Robinson, and many more.

Outside of academia I have been inspired by many individuals whose dedication to their ideals has led them to pursue vocations. I'd like to acknowledge all of my ScriptEd and DS3 students, the ScriptEd volunteers (in particular Andrew Ingraham, an amazingly dedicated individual), and the Hillary Clinton 2016 campaign analytics team.

I did my Ph.D. in a very concrete, non-metaphorical, real world metropolis which wouldn't function without the hard, daily work of a whole lot of people. I'd like to thank some people who helped me get from point A to point B: the workers of the MTA. The subway in NYC isn't in the best shape at the moment but it seems like a pretty thankless and very important job. I'm not sure how many MTA employees will ever see this but... thanks!

# *Dedication*

This work is dedicated to the memory of Tom Hesse.

Chapter 1

# *Introduction*

## 1.1 Motivation

The work presented in this thesis tries to help answer part of a big question: How can businesses, governments, and other organizations collect and use location data without running into problems like over-surveillance or algorithmic bias? Before we dive into this question, however, it's a good idea to step back and think about why this is a problem in the first place. In this introductory chapter, I will break down the question into three areas, defining and examining them in turn:

- **"...organizations collect and use..."** Why are organizations collecting data, and what are they using the data for?
- **"...location data..."** What exactly *is* location data, why is it useful, and what differentiates it from other types of data?
- **"...problems..."** What are the problems that result from organizations using location data?

### Data Collection

Why are businesses, governments, and other organizations collecting and using data? There are many reasons, but perhaps the largest one is simply "money". Organizations can use data to improve their bottom line in many ways, such as finding inefficiencies, discovering emerging business areas, or measuring customer satisfaction. But a very big use for data is ad targeting, with the market size of online advertising estimated to be over $229 billion [139].

The modern day ad targeting industry is the current manifestation of a business model created nearly 200 years ago. The business model is simple and effective: give away a product or entertainment for free and sell the attention or “eyeballs” that go with it. As Tim Wu describes in his book “The Attention Merchants,” [126] in 1833 a man named Benjamin Day founded a newspaper titled “The New York Sun”. The Sun sold for a penny (one sixth the cost of its rivals), losing money on each copy sold but making up for these loses by obtaining higher advertising revenues due to the wide reach of the cheap paper. Physical newspapers eventually gave way to the “free” entertainment of television, about which Andrew Lewis once said “If you are not paying for it, you’re not the customer; you’re the product being sold.” TV in turn gave way to digital sites; now people use social networks and search engines to be entertained and educated, paying not with pennies but with their attention and personal information.

With newspapers, all readers saw the same ads. Of course, the owner of a sports equipment store might pay more to show their ad in the sports section, and likewise a fashionable boutique would pay to have a more prominent ad in the style section. Then, as now, companies paying for advertisements wanted to know that the price they pay is justified. Today, however, it is possible to target ads and display interest (through clicks) at an individual level and there is thus a large drive to pair relevant online ads to the most interested users. Gathering more information about an individual makes it easier to match a purchase interest with a relevant ad in order to generate a potential sale and also to inform companies of the population that are interacting the most with their brands. Some of these matches may strike us as reasonable and useful, such as showing microwave advertisements to an individual searching for information about microwaves. Others may strike us as less ethical, such as services that seek to target customers who are sad or have low self esteem. The fluidity of websites allows such targeting at a fine-grained level, showing ads to users based on the sites they've been to, their suspected interests or incomes, and their locations.

## Location data

The types of information used in targeting models are many. In this section we describe what location data is and why it is a potent ingredient in the modern ad targeting industry. We'll start with how I define location data, how it is collected, and why it is interesting.

In a highly general sense, location data is information relating people to places. Typically, this relation is the fact that a person was at a place at a particular time. However, location data could be information that isn't completely time-bounded: for example, an individual's home address, workplace, or frequent vacation spot. We will typically refer to location data as sets of triples of the form:

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

Where:

$u \in U$ represents a user id from a set of users

$l \in L$ represents a location from a set of locations

$t$ represents time, which may not always be present.

A location $l$ can be described **geographically** or **discretely**. I name data that is defined in terms of coordinates (most typically latitude-longitude) as *geographic* location data. In contrast, *discrete* location data represents locations as a series of IDs or names instead of as points on a plane or sphere. Both of these representations can be connected with *semantic* data, such as a type of venue like restaurant, university, or city.

Location data is generated in multiple ways. Though researchers in other fields may keep track of the locations of animals, robots, or fleets, this thesis focuses squarely on the movement of users, typically consumers. There are two major classes of data capture: **active**, which requires the user to take some action in order for location to be recorded, and **passive**, where a location is recorded periodically without any action from the user. One example of active location capture are "checkins" on location based social networks, where a user pushes a button on an app to tell the world where they are. We will often use the term checkin to denote a single location data point. Another commonly studied form of actively collected location data comes from cell phones via the Call Detail Record or CDR. When cell phone users make a call, the nearest cell tower which services this call is recorded, giving the users location as somewhere within the range of that tower. Note that even though the user did not explicitly try to generate data about their location, it was still captured, and since they took some action to generate the point the data is still labeled active. An example of passively recorded location data would be the constant collecting of Apple and Google's location history or the route taken in the app MapMyRun.

Why is location data useful? Location data reveals lots of information about the preferences and demographics of an individual, key components of a targeted advertising strategy. Knowing where people have shopped in the past, the average home price of the area in which they live, the language typically spoken in their neighborhood, and the method of transportation they use are all pieces of information that can be garnered from location data and used to boost ad revenues. In addition to learning from the previous behaviors of an individual, knowing in real time where someone is can be used by advertisers to push in the moment behavioral suggestions. We show multiple examples of these in this thesis. For instance, in Chapter 3 we show how demographics can be inferred from location data and in Chapter 5 and Chapter 6 we show how location data can be used to boost advertising revenues.

Moving from the ways in which location data can be used into specific methods, we return to the question of location representation. Some typical implementations of machine learning applications handle discretized location better than continuous coordinates, so we will often gather geographic points together into discrete regions, through methods like clustering or truncating coordinate digits. We will refer to the resultant size of the geographic area capturing a set of points as the **granularity** of the data. Likewise, we can use different **time** granularities, such as a minute, hour, or day. How we gather spatiotemporal points together can have a large impact on whatever task we are trying to accomplish. Using very large granularities, where larger regions are gathered into one location, can be more privacy sensitive as they obscures the details or semantic meaning of visits, but by the same token the large granularities might be less effective. Smaller granularities in turn can give more information, narrowing in on cities, to neighborhoods, to a bar on a TV trivia night. The first item in this sequence is clearly very broad whereas the last provides very detailed information on visitor interest. My work experiments with different levels of granularity, showing its impact on, for instance, data uniqueness/anonymity, privacy, and advertising revenue.

There is much more that can be said about the capture, storage, processing, and use of location data. A good technical overview can be found in “Trajectory Data Mining: An Overview” [136].

## Problems

Though location data can be very valuable, the flip side to that coin is location data’s sensitivity. Location data is sensitive because of its ties to physical safety, the difficulty in anonymizing or obfuscating it, and the signal it provides for sensitive attributes.

First and foremost, location data is tied to physical presence and thus is paired with physical safety. Indeed, many of the largest privacy scandals involving the recording or access of location data and have cited concerns around physical safety, such as Google Buzz, [1] [2] Apple, [3] and Uber [4]. The knowledge of someone's location or important places in his or her life like work or home is an important aspect of physical safety.

Beyond physical safety, the collection and use of location data is a concern due to the fact that it is extremely unique and thus is difficult to anonymize or obfuscate. As we discuss in Chapter 2, several years of research by the academic community has shown that most users have location traces that are unique to them and therefore knowledge of just a few of their previous visits can be enough to pull them out of a dataset. Previous research in this area centered around empirical observations about the uniqueness of location data. This research proved the feasibility of attacks where an adversary had access to auxiliary information, or a subset of the "anonymous" data set linked to private user information. Although an issue, these works ignore attacks where a subset of the anonymous data is not available and do not supply a rigorous basis for such attacks. As we show in Chapter 2, data generated from two entirely different domains can be used to link users across data sets through the use of a simple heuristic algorithm.

Removing identifying information connected to location data thus does not safely protect identity. Additionally, humans are very cyclical in their visiting patterns, meaning removing portions of time will also not be very helpful, and our social nature means the movements of others may give away our own. Obfuscation techniques such as reducing granularity, removing certain fields or rows, or using differential

[1] https://www.theregister.co.uk/2010/02/16/google_buzz_security_bug/

[2] https://techcrunch.com/2010/02/12/google-buzz-privacy/

[3] https://arstechnica.com/gadgets/2011/04/how-apple-tracks-your-location-without-your-consent-and-why-it-matters/

[4] http://www.huffingtonpost.com/entry/uber-settlement-god-view_us_568da2a6e4b0c8beacf5a46a

privacy can reduce the utility of the data. Though some loss in accuracy of queries is to be expected, more work needs to be done to rigorously evaluate the costs, both of implementation and in terms of revenue gained or lost. The work presented in Chapter 5 users a revenue model to investigate the impact of a privacy-friendly advertising system.

Beyond simple presence of an individual is what the locations in someone's life may say about him or her. A quote (which I often use) from United States Supreme Court Justice Sonya Sotomayor vividly describes this problem: "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 [118]." At an individual level, control over someone's location history has grave privacy implications. The knowledge of a person's location history can also be the knowledge of their religion, sexuality, habits, strengths, and weaknesses. Such a release can subject people to ridicule, threats, or discrimination: real mental, physical, and economic harms. The prospect of such information being available to others can create a chilling effect on free speech overall. The research community has taken strides towards understanding such threats, but more work remains. What data, and how much data, can be used to infer sensitive attributes? What are attributes for which it is ethical to infer and which are not, and how do we technically capture this when user opinions will drastically defer from one another? Can we measure a "chilling effect" due to data collection or privacy breaches?

These problems do not end just at the level of an individual, but in fact extend to groups. As mentioned previously, location data can be used in machine learning system to target advertisements and improve revenues. By the same token, however, location data correlates with attributes on which we may not wish to target. For example, incorporating an attribute such as ZIP code can in effect create a system that targets based on income or race. Multiple examples of such issues can be found in the popular press [92] as well as the research community. De-biasing our algorithms is an area of research with lots of interest but many open questions. What is the most cost-effective way to de-bias, and will it satisfy either companies or users? For many of these techniques to work, systems will actually need to know the protected classes of the individuals they are trying to aid, requiring *more* data collection, not less. How can we maintain fairness at an individual as well as at a group level? As with almost all of these challenges, can we scale our solutions simply and effectively to the sizes companies now require?

The problems then boil down to this: Economic incentives drive companies to collect large amounts of location data, which is typically done legally and with the consent of their users. Users may not be fully aware of what data is collected, how long it is stored for, or what this data could be used for. Location data is difficult to anonymize, and if an unencrypted version of it is even partially compromised, it can mean harm to users. Additionally, the use of location data in decision-making algorithms can lead to unfairness and discrimination, even if unintentional.

## Prior Solutions

With the use of location data being so desirable but also so potentially fraught with issues, what are we to do?

In a 2008 survey of computational location privacy [60], Krumm presents four main strategies: anonymity, obfuscation, regulatory policies, and privacy policies. The privacy landscape and challenges have also evolved from then, and we might also consider the (often combined) areas of fairness, accountability, and transparency, which could perhaps be thought of as some form of the final three items. We will briefly discuss the list and where opportunities and pitfalls lie.

As we have discussed in the previous section, and as we will demonstrate in Chapter 2, location data is very difficult to anonymize while keeping some unique person-level identifier [132, 83, 80]. Another idea under the banner of anonymity goes further in the direction of user protection: using cryptography to anonymize all location data at a *transactional* level as opposed to at the user level. Blumberg and Eckerlsey [8] outline modern cryptographic techniques which could anonymize location-based search or even automated toll booths, executing the task while retaining no record of who did the searching or paid the toll. For example, some prior work has proposed splitting advertising systems into a local ad service and a cloud-based server, meaning user information can be kept locally and out of the hands of aggregators [45, 40, 115]. Fair billing can be achieved through the use of cryptography. Challenges to their deployment, however, include users who don't fully understand the problem (and so will not advocate for it) and businesses who want to avoid implementation costs and find value in the data beyond the immediate advertisements (such as in product development).

Many papers have explored obfuscation techniques like differential privacy to make concrete the costs of disclosing location data. Though a useful area of exploration, approaches that don't take into account the strong incentives of the actors in the ecosystem will not be successful. Obfuscating or deleting information about users may help companies through avoiding legal costs and bad publicity. However, the preponderance of location data still being collected shows that to most companies the incentives for exploiting this data are currently much higher than the costs. Additionally, statements about probability in terms of data exposure can be difficult for users to interpret, making tools and techniques less likely to generate user enthusiasm or positive press. The work in this thesis tries to deal with these challenges by (1) putting privacy in terms that users can understand, (2) giving users control over their information release while maintaining incentives to continue generating data, and (3) mitigating specific risks of data collection. Rather than obfuscating or anonymizing all data, we provide solutions where data can be collected and stored in a “raw” form.

Branching out from anonymity and obfuscation, the academic community has turned its attention towards to “algorithmic bias” problems by working on solutions for fairness, accountability, and transparency. To be more technical, fairness refers to systems or algorithms that attempt to remove disparity between groups. Accountability focuses on methods to audit or insure compliance with some policy. Transparency tries to take decision processes generated by machine learning and give human-interpretable descriptions of why those decisions took place. Accountability and transparency are key components of regulatory or privacy policies, insuring that companies are actually doing what they propose. Fairness could perhaps be considered a combination of privacy policy and obfuscation. As companies modify their algorithms to insure fairness, they are adhering to a policy through technical means while modifying the way in which user information is used. The later half of my thesis will focus on these ideas of fairness, accountability, and transparency.

## 1.2 Contributions

The work presented in this thesis falls under two main themes. The first portion of the thesis, which includes Chapters 2 and 3, demonstrates empirically new attacks on the anonymity and privacy of users, including a theoretical basis for user identification and expanding our understanding of attacks from individual levels to group levels. In this portion, there are multiple questions we consider. Previous work showed that human mobility is highly unique and hence vulnerable to de-anonymization. What is a reasonable underlying model that could give rise to this phenomenon? Could such a model be used for additional de-anonymization attacks or profile linking? How can we study the cross domain case, when prior research has only considered finding subsets of one dataset? Are there attacks that extend beyond individual identity to sensitive user attributes? What representations of data will be most accommodating to such attacks?

In Chapter 2, we begin by empirically demonstrating that a simple heuristic algorithm can be used to link user identities across "anonymous" datasets using spatiotemporal data. In addition to a test of our algorithm on several novel datasets, the work includes a proof of the effectiveness of our algorithm, based on a model of Poisson-distributed user visits. Prior to the work presented in this chapter, the understanding of uniqueness (and hence, risk of identification) in location data was limited to analysis of single datasets. The only attack vector here was auxiliary information that was a subset of the targeted dataset. In contrast, our work extends to the cross domain case, where datasets generated from entirely different behaviors can be used to re-identify users. Studying the cross-domain case is a difficult task due to requiring two datasets with ground truth links between them, which we overcome by finding novel datasets.

In Chapter 3 we shown that demographics like race and gender can be inferred using only location data. A major challenge in conducting such research is a lack of labeled data, which we overcome through the use of crowd-sourcing and publicly available social network data. We show that such data is representative of the geographies we consider, and explore new metrics of segregation in terms of where people travel as opposed to where they live. In contrast to the previous chapter which focused on anonymity, an individual concern, the possibilities raised in this chapter focus on group concerns.

The second portion of the thesis, Chapters 4, 5, and 6, propose and analyze new solutions for dealing with privacy, anonymity, and fairness issues. Prior work on privacy, while extremely useful, has often presented privacy in an abstract manner that is difficult for average users to understand. For example, differential privacy captures a probability guarantee about whether a user will encounter additional harm from including their data in a database through a parameter $\epsilon$ [26]. $k$-anonymity and the related $t$-closeness and $l$-diversity give facts about the number of similar appearing users in a database, essentially creating plausible deniability [114, 66]. Other solutions propose revealing no user-level information to aggregators. In this section, we work to create privacy solutions that are comprehensible to typical users and economically palatable to aggregators. Conducting such research requires us to ask and answer many questions. How can we inform users with useful and digestable information? How can we give users access to choice and control over their data, while doing so in a way that does not overwhelm them with options? What if we protect against specific harms caused by data aggregation, as opposed to all potential forms of harm? When should solutions be implemented at a local point-of-collection level (such as on a smartphone or in a browser) or at a system level (on an aggregator or advertiser's server)? When should we focus on protecting an individual in contrast to a protected group? Can both be done simultaneously?

In Chapter 4 we work to inform users and provide accountability by creating a personal location privacy auditing tool. In prior chapters, the focus was on concerns of privacy or anonymity. We now shift to work that helps users solve their privacy conundrums. In contrast to tools that cut off access to all user data to advertisers, our goal is to better educate and inform users. The key challenges are (1) to build a functioning, real time site where users can interact with data from multiple sources, and (2) to display information about their locations and privacy in a way that is both informative and additionally easy for users to interpret and understand. The application we build lets users import and visualize the location data collected by popular web services in order to understand what these companies know or can easily infer about them.

In Chapter 5 we design and analyze a system that gives users control over what location data is released to aggregators while preserving the data economy. The key challenge is to devise a way in which users can be properly incentivized to release their data, which we achieve by pricing the data using an auction for digital goods. A second insight exploited for this work is using a representation of locations that are useful to both users and data aggregators. For this, we settle upon "keywords" or semantic representations associated with a location. Furthermore, we use two large scale datasets to analyze the economic impact to advertisers, showing that such a system would not reduce revenues greatly. In our previous work, we provided tools to inform users about privacy. This work, in contrast, takes a more active role, implementing a solution at a system level in order to balance user privacy with the economic incentives to collect data.

In Chapter 6, we conduct a first analysis of the costs to advertisers of implementing "fairness" algorithms in a location-based advertisement setting. Unlike prior works which focus on classifier performance or specific applications, we quantify a revenue-fairness trade-off and use a realistic, sparse dataset of location data. The key challenges to this work are finding an appropriate dataset, implementing a fairness algorithm, and modeling the advertiser revenue. We apply an existing framework for fairness on a dataset gathered from social media, empirically analyzing the potential for inadvertent discrimination among gender and race in location-based systems and additionally showing the impact of location representation on fairness. In contrast to the previous chapter, this work doesn't take aim at the data, but rather at what is *done* with the data by aggregators. Instead of preventing data collection, it analyzes what will happen to revenues when fairness constraints are put into place. Though the work in this chapter is a grounding for research into fairness in location-based ads, our methodology applies to more general advertising tasks.

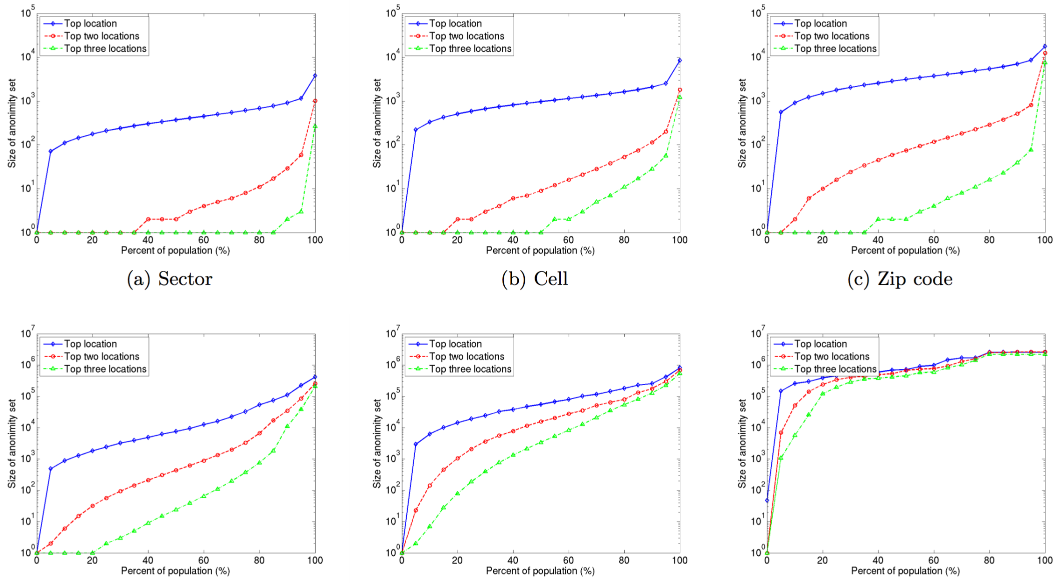
# Chapter 2

# *The Anonymity of Location Data*

When discussing the problems of user data, anonymity is a natural place to start. If it is impossible to link data to a user, how will it be possible to cause that user harm? There are, however, two problems with that question. The first, which we explore in this chapter, is the immense difficulty in anonymizing location data. The second, which we explore later, is that attributes associated with a user can still be used to cause that user harm, regardless of the knowledge of their identity.

The first large-scale study of the $k$-anonymity of location data was appropriately titled "Anonymization of Location Data Does Not Work" [132]. The paper used data from cell phone call detail records (or CDR) for 25 million United States users over a 3 month period. The authors represents each user as simply their top $n$ most visited locations, varying $n$ from 1 to 3. Additionally, the authors varied the granularity of the locations, with the smallest as cell sector and the largest as state. Remarkably, using 3 locations at a cell level made half of all users completely unique, and 3 locations a sector level made 85% of all users unique. A figure detailing this result and results for other granularities and values of $n$ is depicted in Figure 2.1. The authors went on to analyze the impact of geography (comparing different states and cities), mobility (distances between top locations), and social networks on anonymity.

A different study used randomly selected points from a user's dataset and included time of location visit [84], as opposed to a users top $n$ locations (mostly omitting precise time) of Zang and Bolot. Using a call detail record dataset of 1.5 million users from a small European country, this work showed that 95% of users are uniquely



(a) Sector

(b) Cell

(c) Zip code

(d) City

(e) County

(f) State

Figure 2.1: Figure from [132] 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.

identified by 4 spatiotemporal points. A follow up study [81] showed that in a data set of credit card transactions, user profiles of spatiotemporal points had a similar level of uniqueness, and even more when transaction amounts were included as well.

In these works, the question of user anonymity was addressed using either different portions of the same dataset or observing the same behavior across thematically similar domains. In contrast, the general cross-domain case where users have different profiles independently generated from a common but unknown pattern raises new challenges, including difficulties in validation, and remains under-explored. Additionally, previous works primarily showed empirically properties about anonymity in user data. The main contribution of this chapter is a generic and self-tunable algorithm that leverages any pair of sporadic location-based datasets to determine the *most likely* matching between the users it contains. While making very general assumptions on the patterns of mobile users, we show that the maximum weight matching we compute is provably correct. Although true cross-domain datasets are a rarity, our experimental evaluation uses two entirely new data collections, including one we crawled, on an unprecedented scale. The method we design outperforms naive rules and prior heuristics. As it combines both sparse and dense properties of location-based data and accounts for probabilistic dynamics of observation, it can be shown to be robust even when data gets sparse.

The work in this chapter was presented at the World Wide Web conference in 2016 and was conducted with Yunsung Kim, Silvio Lattanzi, Augustin Chaintreau, and Nitish Korula.

## 2.1 Motivation and Summary of Results

Almost every interaction with technology creates digital traces, from the cell tower used to route mobile calls to the vendor recording a credit card transaction; from the photographs we take, to the "status updates" we post online. The idea that these traces can all be merged and connected is both fascinating and unsettling. The ability to merge different datasets across domains can provide individuals with enormous benefits, as seen by increasingly widespread adoption of apps that learn multi-domain user behavior and provide helpful recommendations and suggestions. However, when done by third parties that a user may not interact with directly, this raises fundamental questions about data privacy. In this chapter, we focus on location data and show that this type of data is privacy sensitive. More formally, we focus on the following technical question: Is it possible to link accounts of the same user across datasets using just location data? The answer to that question points both to algorithmic feasibility but also our ability to maintain seemingly distinct identities or personas until one chooses to reveal they belong to the same user.

Increasingly often, as shown in recent studies, the location of a smartphone owner is captured and recorded for a majority of mobile apps even in the absence of geographical personalization. This considerably expands the number of parties who can collect and exploit the knowledge of a user's whereabouts. Even when data is recorded sporadically, these datasets are very rich and intimately connected to one's everyday life; they may present or at least partially reflect our most recognizable patterns. Recently, even a small amount of location information was shown sufficient to either render most users distinguishable [83, 132], or infer multiple sociological traits such as race [103], friendship [18, 21], gender, or marital status when combined with domain semantic information [138].

In spite of this work, determining when and how two accounts belong to the same mobile user in *different* domains remains an open problem, primarily for three reasons: First, identity reconciliation is harder than both classifying and distinguishing users. As an example of the former, one may not be able to connect two profiles exactly, but can still be quite certain that both belong to a high-income American, for instance. For the latter, uniqueness of an individual in one dataset does not imply that they will be easily recognized in another one. For instance, in a simple case where individuals produce location records randomly and independently in two domains, users will likely be unique but it is *provably impossible* to link them across datasets. Second, as a consequence, many previous methods are domain specific and typically focus on clean and dense parts of the data. In contrast, most of our motivating examples above are sparse, and we aim at leveraging locations in the general case without additional information attached. Third, with almost no exceptions, identity reconciliation was always considered for different parts of the exact same data set, or at best domains that are semantically similar. In contrast, our goal is to address the most general case in which records across domains are separately generated but share an underlying pattern: The user's physical location. Since one cannot occupy two locations at the same time, the common pattern of our physical mobility creates fertile ground to notice events that coincide, and those that are incompatible. The main question is how to use those observations (ideally in a provably optimal manner), under which conditions they are sufficient to link accounts, and how to collect data to empirically validate any related claims.

Exploiting rare coincidences to de-anonymize users is now a classic problem, with a sparsity based method available for almost a decade [88]. While we defer a more detailed comparison with our work to the next section, we would like to point out the main ingredient of our algorithm: a new use of misses and repetitions to interpret coincidental records that exploits the sparse property of coupling between Poisson processes. We note that sporadic collection of records typically resembles such statistics for rare events. This method, which is proved optimal and correct under these simple assumptions, is hence particularly effective in various datasets. Another advantage of our scheme is that it relies on only three parameters[1] that are initially unknown but easy to approximate. We prove empirically that simple methods to estimate these parameters are robust even when starting from imperfect observations.

We now present the following contributions.

- A new generic and self-tunable algorithm which combines positive and negative signals from co-incident events to build a new type of maximum weight matching. In practice this algorithm is compatible with a parameter tuning step exploiting a previously proposed density-based method. In spite of no domain-specific tuning, our algorithm outperforms the state of the art.
- A rigorous interpretation of our algorithm justifying its correctness. In particular we provide a simple model of mobility that encompasses various cases of location-based data. This is, to the best of our knowledge, the first mathematical model for observed location traces across multiple domains. We prove the

[1] Two are related, so estimation has two degrees of freedom.

ideal correct matching maximizes our algorithm's score *and conversely*, that only correct matching achieves maximum score in expectation.

- An empirical evaluation of this problem in three distinct scenarios that significantly extends beyond previous studies in both realism and scope. The first dataset, already publicly available, allows immediate comparison with prior results. For the second scenario considered, we collected data from two current live services, gathering considerably more locations, and proving that our method achieves near perfect accuracy. Finally, our method is shown superior in a commercial scenario that is significantly more heterogeneous and challenging[2].

As we explained above, linking anonymous profiles across domains is considerably more challenging than either establishing users' distinguishability or classifying users into different groups. As such, it may have been considered impractical at scale. The fact that we can link users, sometimes with high precision and recall, shines new light on the protection offered by even the most complete anonymity. Our results are, to the best of our knowledge, the first example of a cross domain analysis of this problem to prove an algorithm's correctness, together with the first validation at scale of location based reconciliation in real cases. As more data are available, and different patterns or domain specific properties are discovered, we believe that more algorithms could be designed and evaluated against the technique we present as a benchmark for the most general case.

[2] This dataset was not released in raw form to any researcher in the team; the evaluation was run on a remote server with a non-exclusive agreement that other academic researchers can replicate in the spirit of reproducing and improving future reconciliation methods. Note that the authors from Google did not have even remote access to this data.

## 2.2 Background

It has been shown that most users in location based datasets are unique, either through a few of their most visited places [132] or based on a few timed visits chosen at random [83, 80]. This property follows a tradition of work specifying the risk of releasing even anonymized datasets [114]. What this shows is that users can be re-identified *in theory*, for instance in one of the following two cases: if an adversary has access to auxiliary information (*e.g.*, the real identity of all users who visited a place at a given time, or an original set of seed nodes which are already re-identified) [83], or alternatively if a public data set is known to intersect the anonymized one [114]. What those works do *not* show, however, is how to exploit this uniqueness in the common case we consider: two *distinct* datasets with no auxiliary information that is known *a priori*.

Identity reconciliation so far has leveraged three principles: *Ad-hoc identifying features* such as matching username, email addresses, or unique tags. Those are ignored here; as recently measured in [33] they are rarely available and accurate. *Information propagation*, where starting from a seed set of identified nodes, a graphical structure such as a social network is exploited to expand the set of matched nodes in static [53, 56, 87, 94, 129] or mobile [111, 52] datasets. Again, those techniques cannot be applied in the general case where no preexisting graph and seeds are known[3]. Finally, *identification of nearest neighbors* using similarity metrics [88, 32] generalizes the first method to leverage non-identifying features and imperfect matches. Data sparsity plays an important role, which is typically included in the design of the similarity metric. This approach suffers from the opposite problem: it applies so broadly that it is very loosely defined. Indeed, most successful reconciliations using this tech-

[3] In the most ambitious information propagation where seeds may be noisy and structures, initially unknown, are inferred, the differences between this approach and one based on similarity starts to fade. We experimented with it but found no improvement from information propagation to report.

nique report on the art of deciding upon informative similarity features – or often the subtlety of their combined effects [32] – without necessarily providing a unified justification. Moreover, a closer look showed that the accuracy of similarity methods for static features (e.g., name, home location, friends) are typically overestimated in practice [33]. Our work addresses this important need: Our inference method interprets location datasets, however different in their domains, as sporadic observations of the same hidden mobility processes. We generalize data sparsity from a static viewpoint to a dynamic viewpoint, leveraging naturally misses and repetitions in the observed processes. In spite of a considerable amount of prior work on Entity Resolution [20], we did not find similar analysis and algorithms, probably because mobile datasets are relatively new and exhibit specific dynamics. Similarly, the related literature on network alignment [5] rarely considers the bipartite case [57] and it centers on static graphs. We empirically found that our method yields superior accuracy to those previously proposed, while being more robust and easy to use.

Other attempts at re-identifying users using mobility data *only* have typically expressed similarity between users with *density based methods* [119, 32]. Those rely on a user having a discriminative pattern in the frequency she visits various places. In [119] author aims at reconciling users in the same domain but at different periods, hence ignoring the time of the visits themselves. In situations where datasets overlap in time, those techniques leave much information unused.[4] Another technique, somewhat diametrically opposed, uses *specific visit times* [108]. Prior to this paper, this was only validated in a single domain (by randomly extracting a subset of each user's profile to recognize). We empirically show that none of those methods extend to the more demanding cross domain case without incurring large inaccuracy. This confirms previous observations that density and time based similarities can reduce

[4] It is, for instance, entirely ineffective in a homogeneous population where each user follows the same location distribution for her visits. Our method, in contrast, is proved to correctly handle that case.

the scope of re-identification attacks by removing a lot of dissimilar accounts [32], but cannot be used as is for reconciliation as they lead to low accuracy in practice [33]. Finally, we should mention a statistical learning approach based on Dirichlet distribution used to relate anonymous CDR data with publicly available social network data [13, 14]. It remains, however, difficult to judge its effectiveness as it is used without further theoretical justification and validated without ground truth in the data. Our method, in contrast, is tailored from scratch to location based datasets, its correctness is proved under simple assumption on nodes' visits, and it has been evaluated on three data-sets with ground truth, among the largest available to date, including two that have never appeared in this context. Whether more generic statistical learning reproduces some of the strengths of our method remains an interesting question to explore beyond the scope of this paper.

## 2.3 Location-based Reconciliation

### Problem Formulation and Model

We use $U$ and $V$ to denote the set of $n$ user accounts in the two domains, with accounts to be linked using location-based data. Let $\sigma_I$ denote the true ("identity") mapping that correctly links the two accounts of the each user. The users may visit locations at various times and perform an *action* (such as a checkin), which results in the creation of a record in one of the datasets. Each such record is associated with the location and time-stamp, and possibly additional semantic information that is relevant to this dataset, but may not make sense in a different domain. Therefore, in our algorithm, we *only* use the time-stamped location data. Note that locations and times may be recorded at a different granularity and levels of precision in the two different datasets to be reconciled (for instance, one may only record the nearest cell tower, the other has GPS coordinates). To account for this, we divide locations

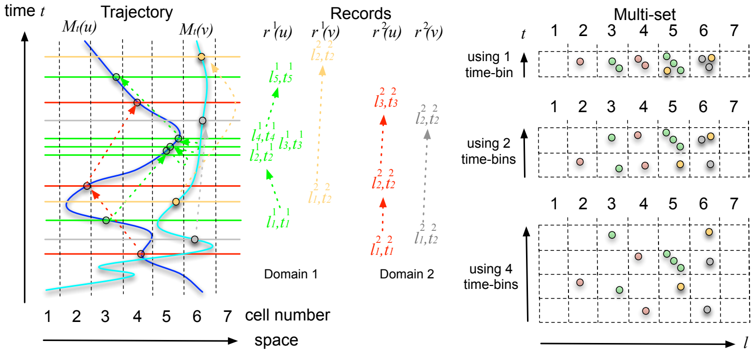


Figure 2.2: Two space-time trajectories with associated footprints in two domains.

and times into *bins*, corresponding to a geographical region or interval of time; For a fixed bin corresponding to location region $\ell$ and time interval $t$, any action recorded in region $\ell$ during time interval $t$ is associated with bin $(\ell, t)$. We use $L$ to denote the set of all location regions and $T$ the set of time intervals in the union of our datasets.

As shown in Figure 2.2, although each user $u$ or $v$ physically follows a continuous time trajectory $M_t$ (shown on the left), her *mobility record* $r(u)$ in each domain is defined as the multi-set of (location, time) bins in which she took an action: $r(u) = \{(\ell_1, t_1), (\ell_2, t_2), \ldots\}$. Note that it is important that this is a multiset: if a user records 2 actions in the same bin, this bin is present twice in the mobility record. Given a specific (location, time) pair $(\ell, t)$ we denote the number of actions in domain 1 that user $u$ took by $a_1(u, \ell, t)$ (i.e., the number of occurrences of $(\ell, t)$ in the multiset $r^1(u)$). We define $a_2(u, \ell, t)$ similarly for domain 2. For ease of notation, we use $a_1$ (respectively $a_2$) to denote $a_1(u, \ell, t)$ (resp. $a_2(u, \ell, t)$) when $u, \ell, t$ are clear from the context.

In this paper, we focus on reconciling users across two domains based only on their mobility records, which we refer to as $r^1(u)$ and $r^2(u)$ respectively. In other words, given a collection of mobility records $\{\ r^1(u) \mid\ u \in U\ \}$ and $\{\ r^2(u) \mid\ u \in V\ \}$ for the same population but with no identity attached, our goal is to return the true mapping $\sigma_I$ which maps the record belonging to one user to the record of the same user in the other collection.

## Mobility Model and Assumptions

In order to formally analyze algorithms applying to the cross-domain reconciliation problem defined above, it is necessary to work under a given *mobility model* which governs how users produce records. Without such assumption, only worst-case performance can be measured, which is arbitrarily bad for any algorithm since one can devise instances where the set of locations with actions in domain 1 is completely disjoint from the set of locations with actions in domain 2. Providing the first such model and proving it leads to a practical method is one of our key contributions.

We assume the mobility records follow a simple generation process: First, for each (location, time) pair, the number of visits of each user to this location during this time period follows a Poisson distribution, with rate parameter $\lambda_{\ell,t}$ and this choice is independent of the visits produced for any other pair. It is a rather crude but effective assumption, as it combines mathematical simplicity (critical later to justify our method), and a form of robustness. Indeed, Poisson distributions are known to be good approximations of rare event processes and to combine gracefully when summed, allowing multiple granularity levels to be combined. They are quite commonly used to handle robust parameter estimation, which is important as the parameter $\lambda_{\ell,t}$ is unknown to the algorithm.

The characterization above describes how visits are produced, but does not specify how users perform actions that are observed. We assume that each time the user visits a location, an action in domain 1 and domain 2 occurs, independently of each other, with probabilities $p_1$ and $p_2$ respectively. Thus, the mobility records are random variables, which we denote by $R^1(u)$ and $R^2(u)$ respectively, with the number of actions in a given bin $(\ell, t)$ being random variables denoted by $A_1(u, \ell, t)$ and $A_2(u, \ell, t)$ respectively. The process of visits and action in each domain is also assumed to be independent among users.

**Possible extensions:** While we keep the model to its simplest form for the sake of a clear exposition, the arguments we provide in this paper generalize to multiple other cases. First among them, all results apply as well when the probability $p_1$ and $p_2$ could depend on $l$ and $t$ as well. One could also analyze our algorithms when those parameters are not constant among users. After experimenting with those more general models, we found that they do not yield significant practical improvement in the scenarios we evaluated. We also note that one can adopt different generative models, but many of these do not change the problem significantly, or the analysis of our algorithm. For instance, the number of visits to a particular location may be generated by a binomial distribution, instead of Poisson.

Other extensions are interesting topics for further study: For example, our model does not currently account for geographical proximity between different locations; in reality, users who visit a location $\ell$ are also likely to visit a nearby location $\ell'$. One advantage is that this keeps our model general and robust to variations in formats and resolution across datasets that are quite common in space-time data. For instance, actions 1km apart may be considered close in a rural setting but far in an urban area. Our method is agnostic to such relative change of distance. We also note that our model ignores dependencies between users. For instance, members of a family may travel together and the presence of friends in a location may render a visit by a given person more likely. On the other hand, our model can accommodate frequency of visits that vary between users and hence create communities that on average visit frequently similar places. With larger and richer data, it is likely that more realistic models than ours may give additional insights and better exploit users' true mobility patterns. However, the simple case we define above leads to a simple algorithm that captures mobility of users sufficiently well to beat the state of the art and present a reasonable benchmark for future use.

## 2.4 Algorithm and Analysis

In this Section, we present an algorithm tailored to the location record model introduced above. Our main contribution is a proof that under these assumptions, there is a tight correspondence between the maximum weight matching that we define and the 'true' matching between users, even exhibiting a positive gap. Later, Section 2.5 will demonstrate that this correspondence generalizes in practice to make this algorithm a superior alternative to multiple known approaches.

### Algorithm

Our algorithm works in two phases: The first phase is to compute a score for every candidate pair of users $(u, v) \in U \times V$ (see below for more details). In a second phase, we first define a complete bipartite graph on $(U, V)$ where the weight of the edge $(u, v)$ is given by the score for $(u, v)$ aforementioned. We then compute the matching in this bipartite graph that has maximum weight[5]. The algorithm then claims that records that are connected by an edge belong to the same user. Under the assumptions introduced above, we can prove that this procedure is always correct.

In the rest of this section, we provide more details on how the scores of a pair $(u, v)$ are determined: For each (location, time) bin $(\ell, t)$, we compute $\mathrm{Score}(u, v, \ell, t) = \ln\left(\phi_{\ell,t}(a_1, a_2)\right)$, where the term $\phi_{\ell,t}$ in the logarithm is:

$$\frac{P\left[A_1(u,\ell,t) = a_1 \wedge A_2(v,\ell,t) = a_2 \mid \sigma_I(u) = v\right]}{P\left[A_1(u,\ell,t) = a_1\right] \cdot P\left[A_2(v,\ell,t) = a_2\right]}.$$

The numerator of $\phi$ measures the probability that the *same* user performs $a_1$ actions in domain 1 and $a_2$ actions in domain 2 in the bin $(\ell, t)$. The two terms in the denominator are the probability that an arbitrary user performs $a_1$ actions in domain 1 in bin $(\ell, t)$, and another user performs $a_2$ actions in domain 2 in this bin. Since

[5] If some edges have negative weight it is possible in theory for a maximum weight matching not to match all users. However, under our assumptions it does not happen.

we assume that user performs actions independently, $\phi_{\ell,t}(a_1, a_2)$ measures how much *more* likely it is to observe $a_1$ actions in domain 1 by account $u$ and $a_2$ actions in domain 2 by account $v$ if these accounts belong to the *same user* than if these are two different users.

Note that, in the above definition of $\phi_{\ell,t}$, the probability is taken in the model we introduce (*i.e.*, that of independent actions taken conditioned on Poisson visits). This yields multiple equivalent formulas to compute the ratio $\phi_{\ell,t}$:

**Lemma 1.** *The value of $\phi_{\ell,t}(a_1, a_2)$ in the model we introduce is equal to any of the following expressions (where $\lambda_{\ell,t}$ is denoted by $\lambda$ for ease of notation):*

*(i)* $$\frac{P\left[A_1(u,\ell,t) = a_1 \wedge A_2(v,\ell,t) = a_2 \mid \sigma_I(u) = v\right]}{P\left[A_1(u,\ell,t) = a_1\right] \cdot P\left[A_2(v,\ell,t) = a_2\right]}.$$

*(ii)* $$\frac{e^{-\lambda} \sum_{k \geq \max(a_1,a_2)} \frac{\lambda^k \binom{k}{a_1}(1-p_1)^{k-a_1} \binom{k}{a_2}(1-p_2)^{k-a_2}}{k!}}{\sum_{k \geq a_1} \frac{\lambda^k \binom{k}{a_1}(1-p_1)^{k-a_1}}{k!} \cdot \sum_{k \geq a_2} \frac{\lambda^k \binom{k}{a_2}(1-p_2)^{k-a_2}}{k!}}.$$

*(iii)* $$\frac{e^{-\lambda(1-p_1-p_2)}}{(\lambda(1-p_1))^{a_1}(\lambda(1-p_2))^{a_2}} \sum_{k \geq \max(a_1,a_2)} \frac{(\lambda(1-p_1)(1-p_2))^k k!}{(k-a_1)!(k-a_2)!}.$$

*(iv)* $$\frac{e^{-(\lambda p_1 p_2)}(1-p_1)^{a_2}(1-p_2)^{a_1}}{(\lambda(1-p_1)(1-p_2))^{\min(a_1,a_2)}} \mathbb{E}\left[\frac{(X+\max(a_1,a_2))!}{(X+|a_1-a_2|)!}\right],$$

*for expectation taken over $X$ a Poisson variable with parameter $r = \lambda(1 - p_1)(1 - p_2)$.*

*Proof.* $(i)$ becomes $(ii)$ once we develop each probability by conditioning on the number of visits $k$ that $u$ and/or $v$ make to the bin $(\ell, t)$, and we observe that a few terms simplify. To obtain $(iii)$ one should observe by the Poisson sampling property that $A_1(u, \ell, t)$ is also distributed according to a Poisson variable, with parameter $(\lambda p_1)$. This simplifies the denominator which then yields this expression. Finally, to obtain $(iv)$, it suffices to introduce the change of variables $k' = k - \max(a_1, a_2)$ and notice that the series becomes this expectation taken over all possible values taken by $X$. □

Our algorithm, formalized immediately below, can leverage any of the above formulas to approximate $\phi$. Expression $(i)$ is the most general (and holds even for non-Poisson visits). Using $(iv)$ with $p_1 = p_2$ and $a_1 = a_2 = a$ we see that the score is especially large when $\lambda$ is small (as this visit is rare) and $a$ is large (the common observations occurs more than once). For each pair of records, the algorithm computes all the scores associated with the (location,time) bins. It sums them across all bins to compute the weight of the edge between this pair.

**Algorithm 1** Our reconciliation algorithm

```
**Require:** $\forall u \in U : r^1(u), \forall v \in V : r^2(v), \{\lambda_{\ell,t}\}$
  **for** $(u,v) \in (U \times V)$ **do**
    $w(u,v) = \sum_{t\in T} \sum_{\ell \in L} \ln \phi_{\ell,t}\left(a_1(u,\ell,t), a_2(v,\ell,t)\right)$
  **end for**
  Let $E = \{w(u,v) : (u,v) \in (U \times V)\}$
  Compute the maximum weighted matching on the bipartite graph $B(U,V,E)$
  **return** the function that maps matched vertices.
```

While the algorithm is conceptually well defined, there are two things to note about its implementation. First, the input includes the set of parameters of the Poisson distribution, $\{\lambda_{\ell,t}\}$; these are not known, but can be estimated (see discussion in Section 2.5). Second, the definition of $\phi$ involves infinite sums over all values of $k \geq a_1, a_2$. We prove below that this can be approximated to arbitrary precision by taking the sum over a limited number of terms.

We now justify our algorithmic approach, and prove that the expected score is highest for the true matching.

## Relation to Maximum Likelihood

We explain our choice of the function $\phi$ (and hence our specific weight function $w(u,v)$) by showing that the weight of a matching is proportional to its log likelihood, and the matching with maximum expected weight (i.e. maximum expected likelihood) is indeed the true matching $\sigma_I$.

The observed inputs to the algorithm are the mobility records $r^1, r^2$. Taking a maximum likelihood estimation (MLE) approach, our goal is to find the matching or permutation $\sigma$ that maximizes the likelihood $P[\sigma \mid r^1, r^2]$. As is standard, we have:

$$P[\sigma \mid r^1, r^2] = \frac{P[R^1 = r^1, R^2 = r^2 \mid \sigma] \cdot P[\sigma]}{P[R^1 = r^1, R^2 = r^2]}$$

Assuming a uniform prior over all permutations $\sigma$, it is easy to see that we are trying to find the permutation $\sigma$ maximizing $P[R^1 = r^1, R^2 = r^2 \mid \sigma]$.

Assuming $\sigma$ is the true permutation / mapping, since mobility of different users is independent, the probability of observing various actions for $u$ depends only on the actions of $\sigma(u) = v$. Therefore, we have:

$$P[R^1 = r^1, R^2 = r^2 \mid \sigma] = \prod_{u,v:\sigma(u)=v} \prod_{\ell \in L} \prod_{t \in T} P[a_1(u, \ell, t), a_2(v, \ell, t) \mid \sigma_I(u) = v] \quad (2.1)$$

To normalize this probability, we divide by the overall probability of observing $r^1$ and $r^2$ in the two domains. Since $P[R^1 = r^1] = \prod_u \prod_{(\ell,t) \in L \times T} P[A_1(u, \ell, t) = a_1(u, \ell, t)]$ and $P[R^2 = r^2] = \prod_v \prod_{(\ell,t) \in L \times T} P[A_2(v, \ell, t) = a_2(v, \ell, t)]$ we note in particular that $P[R^1 = r^1] \cdot P[R^2 = r^2]$ does not depend on $\sigma$. Hence dividing Eq.(2.1) by it does not change which $\sigma$ maximizes the likelihood.

Combining these, it is easy to observe that the likelihood of $\sigma$ is proportional to:

$$\frac{P[R^1 = r^1, R^2 = r^2 \mid \sigma]}{P[R^1 = r^1] \cdot P[R^2 = r^2]} = \prod_{u,v:\sigma(u)=v} \prod_{(\ell,t) \in L \times T} \phi_{\ell,t}(a_1(u, \ell, t), a_2(v, \ell, t)$$

Taking the logarithm of both sides, we see that the log likelihood is proportional to:

$$\sum_{u,v:\sigma(u)=v} \sum_{(\ell,t) \in L \times T} \ln \phi_{\ell,t}(a_1(u, \ell, t), a_2(v, \ell, t)) = \sum_{u,v:\sigma(u)=v} w(u, v)$$

To put it differently, this proves that the log likelihood of $\sigma$ is exactly the weight of the matching it defines in the bipartite graphs that our algorithms constructs. Hence, constructing a maximum-weight matching as our algorithm does is equivalent to computing the maximum-likelihood permutation $\sigma$ given our observations.

What remains to be shown is that maximum likelihood exhibits a gap, *i.e.*, the correct permutation $\sigma_I$ reconciling identity of all users has an expected weight that is higher than any other permutation by a positive margin. Note that, since $\phi$ involves infinite sums, we need to prove this result for the approximated expected weight that we obtain after truncating each sum in the definition of $\phi$.

## Proof of Correctness

Recall that for each location $\ell$ and time $t$, we compute a score for a pair of users $u$ and $v$ based on the number of observed actions $a_1(u, \ell, t)$ and $a_2(v, \ell, t)$ as the logarithm of the function $\phi_{\ell,t}$. Fixing $\ell, t$, we drop the subscripts and simply write $\lambda = \lambda_{\ell,t}$ and $\phi = \phi_{\ell,t}$. We defined $\phi(a_1, a_2)$ as:

$$\frac{e^{\lambda} \sum_{k \geq \max\{a_1,a_2\}} \frac{\lambda^k}{k!} \binom{k}{a_1} (1-p_1)^{k-a_1} \binom{k}{a_2} (1-p_2)^{k-a_2}}{\sum_{k \geq a_1} \frac{\lambda^k}{k!} \binom{k}{a_1} (1-p_1)^{k-a_1} \cdot \sum_{k \geq a_2} \frac{\lambda^k}{k!} \binom{k}{a_2} (1-p_2)^{k-a_2}}$$

Note that this requires taking three infinite sums, but to define a practical algorithm, we cannot sum over an infinite number of terms. We now argue that for any $C$, we can efficiently approximate $\phi$ to within $\pm 1/C$. More formally

**Theorem 1.** *Let $C \geq e^7$ and $\phi'(a_1, a_2)$ be defined using the above definition of $\phi(a_1, a_2)$ by truncating the numerator after* $\max\{\ln C, 2\max\{a_1, a_2\}\}$ *terms, and each factor in the denominator after* $\ln C$ *terms. We then have*

$$1 - \tfrac{1}{C} \leq \tfrac{\phi'(a_1,a_2)}{\phi(a_1,a_2)} \leq 1 + \tfrac{1}{C}\,.$$

We now show that the expected weight of the true / identity permutation is larger than the expected likelihood of any other permutation by a constant, even after truncating the calculation of $\phi(a_1, a_2)$.

**Lemma 2.** *For any bin $(\ell, t)$ and any pair of users $(u, v)$, then $v \neq \sigma_I(u)$ implies $E[Score(u, v, \ell, t)] \leq 0$. On the other hand, $v = \sigma_I(u)$ implies $E[Score(u, v, \ell, t)] > \lambda_{\ell,t} p_1^2 p_2^2 K$, where $K = \frac{1}{2}\lambda(p_1 + p_2 - p_1 p_2)^2$.*

*Proof.* Since we have a fixed $\ell, t$, we use $\phi$ to denote $\phi_{\ell,t}$, $\lambda$ to denote $\lambda_{\ell,t}$, and $A_1(u), A_2(v)$ to denote $A_1(u, \ell, t)$ and $A_2(v, \ell, t)$ respectively. First, consider the case $v \neq \sigma_I(u)$. The expected value of $\phi$, *i.e.*, $E[\phi(A_1(u), A_2(v))]$ can be rewritten:

$$\begin{aligned}
&\sum_{a_1,a_2} P[A_1(u) = a_1]P[A_2(v) = a_2] \cdot \phi(a_1, a_2) \\
= &\sum_{a_1,a_2} P[A_1(u) = a_1]P[A_2(v) = a_2] \\
&\quad \times \left(\frac{P[A_1(u) = a_1 \wedge A_2(v) = a_2 \mid v = \sigma_I(u)]}{P[A_1(u) = a_1] \cdot P[A_2(v) = a_2]}\right) \\
= &\sum_{a_1,a_2} P\left[A_1(u) = a_1 \wedge A_2(v) = a_2 \mid v = \sigma_I(u)\right] = 1
\end{aligned}$$

where the final equality comes from summing probabilities over the entire domain of the joint distribution. By Jensen's inequality:

$$\begin{aligned}
E[\text{Score}(u, v, \ell, t)] &= E[\ln \phi(A_1(u), A_2(v))] \\
&\leq \ln E[\phi(A_1(u), A_2(v))] = \ln 1 = 0
\end{aligned}$$

We now consider the harder case, when $v = \sigma_I(u)$.

$$\begin{aligned}
E[\text{Score}(u, v, \ell, t)] &= E[\ln \phi(A_1(u), A_2(v))] \\
&= \sum_{a_1,a_2} P[A_1(u) = a_1 \wedge A_2(v) = a_2 \mid v = \sigma_I(u)] \cdot \ln \phi(a_1, a_2).
\end{aligned}$$

To simplify notation below, we use $X(a_1, a_2)$ to denote $P[A_1(u) = a_1 \wedge A_2(v) = a_2 \mid v = \sigma_I(u)]$, and $Y(a_1, a_2)$ to denote $P[A_1(u) = a_1] \cdot P[A_2(v) = a_2]$. The distributions $X$ and $Y$ give the probabilities of observing $a_1$ and $a_2$ actions in the two domains assuming the users are the same, and are not the same respectively. Using this notation, we have:

$$E[\text{Score}(u, v, \ell, t)] = \sum_{a_1,a_2} X(a_1, a_2) \ln \frac{X(a_1, a_2)}{Y(a_1, a_2)} = I(A_1; A_2)$$

where $I(A_1; A_2)$ denotes the mutual information between $A_1$ and $A_2$, which is also equal to $D_{KL}(X \parallel Y)$, the Kullback-Leibler (KL) divergence of $Y$ from $X$; this quantity is always non-negative.

We have already shown that for $v \neq \sigma(u)$, the expected score is at most 0. On the other hand, for $v = \sigma(u)$, we have the expected score being non-negative. However, we wish to go further and prove that $E[\text{Score}(u, v, \ell, t)]$ is lower bounded by a positive constant in the latter case.

To do this, we apply the following lower bound:

$$\begin{aligned} I(A_1; A_2) &= X(0,0) \ln \frac{X(0,0)}{Y(0,0)} + \sum_{a_1, a_2 \neq (0,0)} X(a_1, a_2) \ln \frac{X(a_1, a_2)}{Y(a_1, a_2)} \\ &\geq X(0,0) \ln \frac{X(0,0)}{Y(0,0)} + (1 - X(0,0)) \ln \frac{(1 - X(0,0))}{(1 - Y(0,0))} \,. \end{aligned}$$

We now evaluate $X(0,0)$ and $Y(0,0)$ respectively.

$$\begin{aligned} X(0,0) &= \sum_{k \geq 0} e^{-\lambda} \frac{\lambda^k}{k!} (1 - p_1)^k (1 - p_2)^k \\ &= e^{-\lambda(p_1 + p_2 - p_1 p_2)} \sum_{k \geq 0} e^{-\lambda(1 - p_1)(1 - p_2)} \frac{(\lambda(1 - p_1)(1 - p_2))^k}{k!} \\ &= e^{-\lambda(p_1 + p_2 - p1 p_2)} \geq 1 - \lambda(p_1 + p_2 - p_1 p_2) \end{aligned}$$

where the last equality is because the preceding sum contains all probabilities from a Poisson distribution with rate parameter $\lambda(1 - p_1)(1 - p_2)$, and the final inequality comes from the Taylor series expansion of $e^{-x}$. Similarly, we have:

$$\begin{aligned} Y(0,0) &= \left( \sum_{k \geq 0} e^{-\lambda} \frac{\lambda^k}{k!} (1 - p_1)^k \right) \cdot \left( \sum_{k \geq 0} e^{-\lambda} \frac{\lambda^k}{k!} (1 - p_2)^k \right) \\ &= e^{-\lambda p_1} e^{-\lambda p_2} \\ &= e^{-\lambda(p_1 + p_2)} \end{aligned}$$

This yield a lower bound on the mutual information above:

$$\begin{aligned} \text{First,} \quad & X(0,0) \ln \frac{X(0,0)}{Y(0,0)} \\ & \geq (1 - \lambda(p_1 + p_2 - p_1 p_2)) \ln \frac{e^{-\lambda(p_1 + p_2 - p_1 p_2)}}{e^{-\lambda(p_1 + p_2)}} \\ & = (1 - \lambda(p_1 + p_2 - p_1 p_2)) \lambda p_1 p_2 \,. \\ \text{Then} \quad & (1 - X(0,0)) \ln \frac{(1 - X(0,0))}{(1 - Y(0,0))} \\ & \geq \lambda(p_1 + p_2 - p_1 p_2) \ln \frac{(1 - e^{-\lambda(p_1 + p_2 - p_1 p_2)})}{(1 - e^{-\lambda(p1 + p_2)})} \end{aligned}$$

| Dataset | Domain | Users | Checkins | Median<br>Checkins | 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 Uses | 452 | ∼40k | ∼60 | ∼3500 | 2013-04 – 2013-07 |

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

Combining these terms and applying algebraic manipulation yields the desired result with the appropriate value of $K$. This work was primarily completed by my collaborators, Silvio Lattanzi and Nitish Korula, though with the assistance of the other authors (Yunsung Kim, Augustin Chaintreau, and myself). As such, I place this proof in the appendix (see Appendix A) instead of here. □

## 2.5 Comparison and Case Studies

Having established the theoretical guarantees for our algorithm, we now compare its performance to alternative reconciliation algorithms, inspired by the state of the art. We describe our datasets, the baselines we compared against, some of our real-world implementation, and our results.

### Datasets

Studying the cross domain problem is challenging due to the difficulty in obtaining ground truth. We used a total of three datasets (each from different pairs of spatio-temporal domains) to evaluate the performance of Algorithm 1.

**Foursquare–Twitter** Our first dataset, labeled **FSQ-TWT**, links checkins on the location-based social network, Foursquare, to geolocated tweets. This dataset was collected previously in [135]. After selecting users with locations present in both dataset, we obtain 862 users with 13,177 Foursquare checkins and 174,618 Twitter checkins.

This dataset presents an interesting challenge. There is a large imbalance in data, with many more tweets than Foursquare checkins.

Additionally, the domains are somewhat different– whereas Foursquare checkins are typically associated with a user showing what they are currently doing (in particular, eating at a restaurant), tweets are more general and associated with more behaviors. To verify that tweets and checkins were usually not one event forwarded by software across both services, which could make this dataset artificially easy, we looked at if checkins matched *exactly* on time place. Only 260 pairs of checkins (less than 0.3%) had exactly matching GPS coordinates, and of those, none were within 10 seconds of each other. Beyond this, we reduced all coordinates to 4 digits of accuracy (around 10m), removing low level GPS digits that could be used as a "signature".

**Instagram–Twitter** Our second dataset, referred to as **IG-TWT**, links users on the photo-sharing site, Instagram, to the microblogging service, Twitter. We obtained this data in the following manner: First, we download publicly available location data from Instagram, saving user metadata if he or she had at least 5 geotagged photos in their 100 most recently uploaded photos. For each photo, we did not download or save any images, instead only using latitude-longitude pairs, times, and a user identifier. To find more profile IDs to crawl, we used the profile IDs of anyone who commented or "liked" a crawled user's photos. We started this process with the founder of Instagram, a central node whose photos are commented on or receive "like" from a diverse set of users. This process yielded 120K users with 35M checkins (i.e. time, latitude-longitude pairs from a geolocated photo).

On Instagram, a user can associate a single URL with their profile. We analyzed these URLs, looking for URLs which matched Twitter accounts. Of these, we manually examined 50, finding that all profiles were correct matches based on profile name, profile picture, and/or posted photos, when available. Then, using Twitter's API, we crawled all publicly available tweets for those users, again saving latitude-longitude pairs, time, and user identifier for geolocated tweets. This process left us with 1717 matched users, with a total of 337,934 Instagram checkins and 447,366 Twitter checkins.

This dataset promises to be the "easiest", due to the large number of photos and tweets per user (median 93 and 89, respectively). Picture-taking and tweeting appear to be somewhat different behaviors, but related in the sense that both are actions whereby a user communicates an action or message to a larger, public audience. To again verify that tweets and Instagram posts were not one event forwarded to both services via software, we again looked at exact matches in low-level GPS coordinates and time. Only 2415 pairs of checkins (around 0.6% of all checkins) had exactly matching GPS coordinates, and of those, only 2 were within 10 seconds of each other. Again, all coordinates were then reduced to 4 digits.

**Cell Phone – Credit Card Record** Our third and final dataset contains a log of phone calls (referred to as call detail records or "CDR") linked to credit card transactions (referred to as "bank" data) made by 452 users from a G20 country over 4 months from April 1st through July 31st, 2013. We will refer to this dataset as **Call-Bank**. The linking was made by two companies who originated the data, a telecommunications and credit card company, respectively. Each record of a phone call in the CDR data consisted of a phone number, time, and cell tower ID with its latitude-longitude coordinates. Each record of a credit card transaction in the bank data consisted of the latitude and longitude of the geolocated business at which the transaction was made, along with the time and phone number of the credit card owner. These transactions only included in-person visits, as opposed to online or over-the-phone transactions. The two companies hashed the phone number using the same hash function, and associated this hash with the information for that user. This information was then passed to a third party. The researchers from Columbia University accessed this information on a secure, remote server.[6] At no time were the real phone numbers or credit card numbers available or utilized.

The two datasets log location in different ways. For the CDR data, a user could have been anywhere within range of the associated cell tower. The bank data, however, have a more precise localization. To link the two, we compute the Voronoi diagram generated by cells' locations. We then say that a business location is the same as a cell tower if it is contained in this tower's Voronoi cell. Note that this is a clear demonstration of the need for location *bins* (in this case, the Voronoi cells), as introduced in our model.

The original data is extremely sparse, and contains above 70k users common to the two datasets. However, many users have no calls or bank transactions in the same location, because about 80% of users have fewer than 10 transactions, meaning they use their credit card on average roughly once every two weeks. To make the problem more tractable, we used a smaller subset of active users, by discarding those that made fewer than 50 bank transactions throughout the entire span (*i.e.*, keeping those making a transaction on average every 2-3 days). It amounts to a total of 452 users, whose transactions and calls are dispersed throughout a total of over 3500 cell towers.

This dataset promises to be extremely challenging. Phone calls and credit card transactions are very different activities, and it is not expected that they occur for

[6] The researchers from Google never had access to this data.

a user in the same place at the same time. Indeed, only 294 of our 452 active users had even at least one location in common across domains.

**Summary** We summarize the statistics on the datasets in Table 2.1. Note that although our datasets have the same set of users in both domains, our algorithm can run *without* this requirement– our algorithm will simply leave some users unmatched. Although by some standards these datasets are small, their size is comparable to previous studies [135, 108] and it is difficult to obtain cross-domain datasets of greater magnitude while still maintaining high levels of accuracy.

## Prior Algorithms

We compare our algorithm with three state of the art reconciliation techniques, which we briefly describe in the rest of this subsection.

**Exploiting Sparsity: The "Netflix Attack"** The first reconciliation technique that we consider is a variation of the algorithm used to de-anonymize the Netflix prize dataset [88]. The Netflix algorithm cannot be applied directly to our setting, but is not hard to adapt. The algorithm first defines a score between users $u$ and $v$ as follows:

$$S(r^1(u), r^2(v)) = \sum_{(l,t) \in r^1(u) \cap r^2(v)} w_l f_l(r^1(u), r^2(v)) \,,$$

where $w_l = \frac{1}{\ln\left(\sum_{v,t} a_2(v,l,t)\right)}$ and $f_l(r^1, r^2)$ is given by

$$e^{\frac{\sum_t a_1(u,l,t)}{n_0}} + e^{-\frac{1}{\sum_t a_1(u,l,t)} \sum_{t:(l,t)\in r^1} \min_{t':(l,t')\in r^2} \frac{|t-t'|}{\tau_0}} \,.$$

Note that $n_0$ and $\tau_0$ are unspecified parameters of the algorithms. This score function considers the visits of $u$ to the locations near $v$'s trajectories. In resemblance to the score function in [88], it favors locations that are visited less often, as they are considered more discriminative just like in [32], frequent visits to the same location, and visits that occur shortly before or after $v$'s traces. The algorithm declares a user $u$ with the best score to be a match for a user $v$ if the score of the best candidate and the score of the second best candidate differ by no less than $\varepsilon$ standard deviations of all candidate scores - otherwise the user is unmatched. Intuitively, this algorithm is designed to exploit *sparsity*, using unique, rare occurrences in two datasets to link users. For future use, we refer to this algorithm as NFLX.

**Exploiting Density: Histogram Matching** In [119] the authors leverage frequency of visits to location as a fingerprint of individuals across datasets. Let $\Gamma^1_l(u)$ be the fraction of time that user $u$ is in location $l$ in the first dataset and $\Gamma^1(u)$ be the distribution across different locations. For each pair of user $u$ and $v$ the weight $w(u,v)$ between them is defined using the Kullback-Leibler divergence:

$$D\left(\Gamma^1(u)\,\middle\|\,\frac{\Gamma^1(u)+\Gamma^2(v)}{2}\right)+D\left(\Gamma^2(v)\,\middle\|\,\frac{\Gamma^2(v)+\Gamma^1(u)}{2}\right).$$

Each edge weight reflects the degree of disparity between two users. This algorithm computes a minimum weight matching for the complete bipartite graph drawn between individuals, as a way to minimize that disparity. In contrast to NFLX, this algorithm relies on the *density* of data, assuming that over time even in different periods a unique histogram of user visits will emerge from a user's behavior. In the remaining we refer to this technique as HIST. Note that other methods use frequency of visits to define similarity, such as [32]. It can be shown under similar assumptions to our model that within the categories of algorithms that only leveraging density, HIST provably provides the minimum error and that it decreases fast as more data are available [120].

**Alternative: Frequency-Based Likelihood** As a third comparison we consider the reconciliation technique introduced in [108], which approximates the likelihood of a visit made in one domain by the frequency of visits for that user in the other domain, hence assuming:

$$P\left(l \mid r^1(u)\right) = \frac{\sum_t a_1(u,l,t) + \alpha}{\sum_{l',t} a_1(u,l',t) + \alpha|L|}\,,$$

where $\alpha > 0$ is a parameter. This regularization, sometimes referred to as Laplacian smoothing, prevents null empirical frequencies from leading to an infinite score. The mapping (that we denote by WYCI after the title of the paper) is then computed as $\sigma(u) = \arg\max_v \prod_{(l,t)\in r^2(v)} P\left(l \mid r^1(u)\right)$. The paper introduces another distance parameter, but later claims it has negligible impact, as we also observe ourselves.

## Implementing Algorithm 1 in Practice

**Parameter Estimation** In our experiments we partition the time interval into 1024, 2048, 3072 and 4096 time bins. In each time bin we de-duplicate visits to the same locations. In the rest of the paper we describe the results for 4096 time bins, although as we show, similar results hold for different binning.

Our algorithm requires knowing the three main parameters $p_1, p_2$ and $\lambda_{l,t}$ for each bin $(l,t)$. Unfortunately, using single domain observations separately, the problem is ill posed. For instance parameters $(p_1, p_2, \lambda)$ and $(\frac{p_1}{2}, \frac{p_2}{2}, 2\lambda)$ are simply indistinguishable from a marginal standpoint. On the other hand, by conditioning on bins $(l,t)$ where an action in domain 1 is observed, we have

$$p_2 \approx \frac{\sum_u \sum_t \sum_l \min(a_1(u,l,t), a_2(\sigma_I(u),l,t))}{\sum_u \sum_t \sum_l a_1(u,l,t)}\,,$$

at least in expectation. But this formula requires knowing $\sigma_I$, which is precisely the unknown we aim to find. A critical observation we make is that approximating $p_1$ and $p_2$ is good enough. All we need is a candidate permutation $\sigma$ to match user across different domains only for the sake of parameter estimation. In our experiment we use the output of the HIST as our candidate permutation $\sigma$. While it is possible to iterate once a new permutation is found to refine even further, we observe in practice that it is not necessary.

Finally, we have to estimate $\lambda_{l,t}$. Unfortunately most datasets are sparse and do not allow separate estimation of $\lambda_{l,t}$ accurately at each time and location. However, we found that assuming that $\lambda_{l,t}$ is constant across time allows a first estimate of a location-normalized popularity given by $\rho_l \approx \frac{\sum_u \sum_t a_i(u,l,t)}{\sum_u \sum_t \sum_l a_i(u,l,t)}$. The parameter $\lambda$ can then be computed by aggregating observations on all locations together with normalizing factors removed:

$$\lambda \approx \frac{1}{(|U|+|V|)|T|} \sum_l \left( \frac{\sum_{u,t} a_1(u,l,t)}{p_1 \rho_l} + \frac{\sum_{v,t} a_2(v,l,t)}{p_2 \rho_l} \right).$$

Later, we show that estimated parameters are quite robust and resemble ground truth estimated from the true matching.

**Additional Feature** Finally, we introduce for practical settings an "eccentricity" factor $\varepsilon$, which works as follows. After a matching is computed, we only output this edge if the matched candidate's score differs from the second-best by more than $\varepsilon$ times the standard deviation of all candidates.

## Comparison on Real Cases

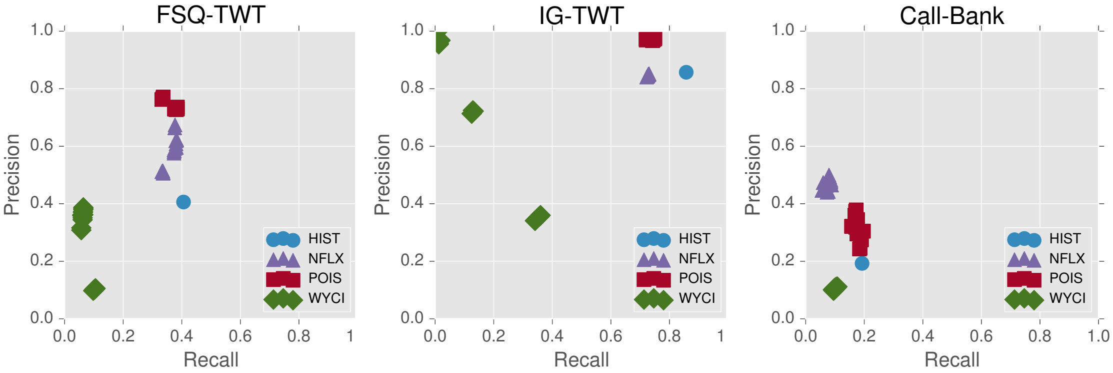


Figure 2.3: Precision and Recall plots for each dataset.

We now turn our attention to experimental performances of our algorithm. In Figure 2.3, we show the precision recall plots for our algorithm (for different eccentricity values) and for the other three reconciliation techniques: HIST, NFLX and WYCI. For our algorithm, we used estimated parameters and for the other techniques, we used optimal parameters (found via exhaustive search).

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

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

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

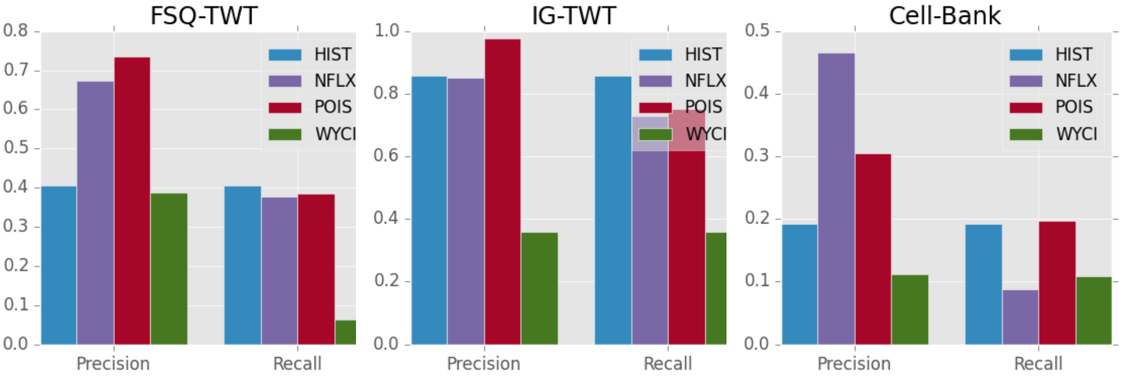


Figure 2.4: Best precision and recall performance for each technique in various datasets.

In Figure 2.4 we present the best performances of the four techniques in the three dataset. It is interesting to notice that our algorithm gives the best trade-off between precision and recall. In particular, even if other techniques achieve sometimes better precision or recall our algorithm is not dominated by other algorithms. In fact it is always Pareto optimal in respect of the precision recall curve, and the only algorithm for which this is true.

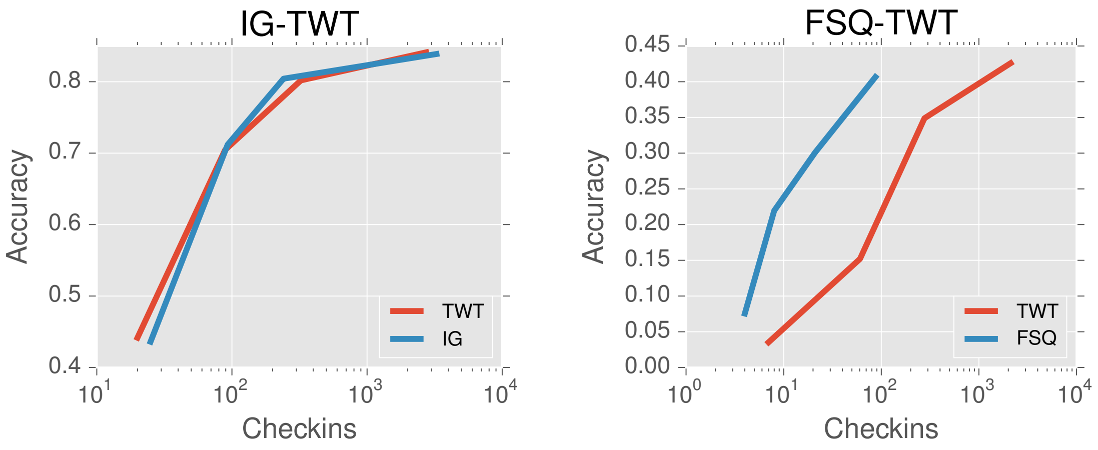


Figure 2.5: Number of checkins vs. our algorithm's accuracy.

We now investigate the impact of the number of user checkins on accuracy. In Figure 2.5, by binning users into quartiles based on number of checkins, and observing the accuracy, we can see that that our algorithm is able to leverage both the amount of the data and its uniqueness. In fact the performance of our algorithm are positively correlated both with the number of checkins and with the entropy of the visited location.

We next turn our attention to the impact of our estimated parameters. As mentioned in Sec. 2.5, we cannot know the exact values of $p_1, p_2$, and $\lambda_{l,t}$. When running our algorithm, we first found a guess at a permutation, and used that matching to estimate the parameters. Comparing this with using the *true* permutation, we can see how far off our guess was and the impact on the algorithm. Fig. 2.6 shows two lines, one using parameters derived from the real permutation and one using an estimate. Clearly, using the estimate is as good as using the real permutation, and is in fact better at certain time levels. Additionally, this figure shows that there is only a small

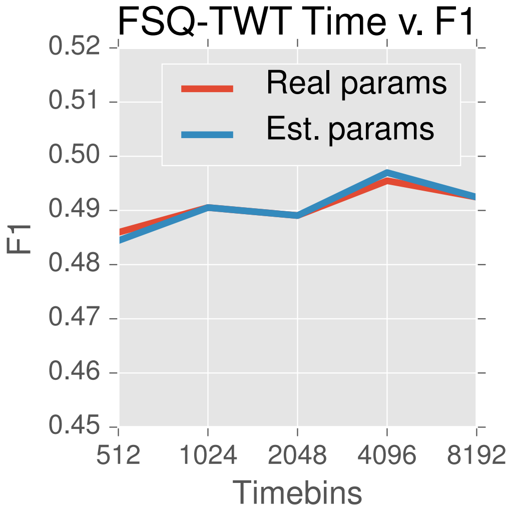


Figure 2.6: Effect of parameter estimation and time binning on algorithm performance.

boost in performance when using differently sized time bins. This is helpful in that it seems the algorithms performance is largely unaffected by choice of parameters.

Finally we show in Figure 2.7 the effect of eccentricity and number of terms (of the infinite sum) on performances of our algorithm. The eccentricity is a term that rejects links if other candidates are also very likely. A higher eccentricity should thus correspond with greater precision at the cost of lower recall. In these figures, we can see that this relationship indeed holds, allowing users to potentially find only the strongest matches, perhaps as “seed” links for other algorithms. The number of terms appears to have little effect on algorithm performance, empirically validating our proof that our approximation appears to have little impact on the final result.

## 2.6 Conclusion

User data is constantly multiplying across an increasing array of websites, apps and services, as they are eager to share part of their behavior with service providers to receive personalized (and free) services. Users may attempt to deal with the

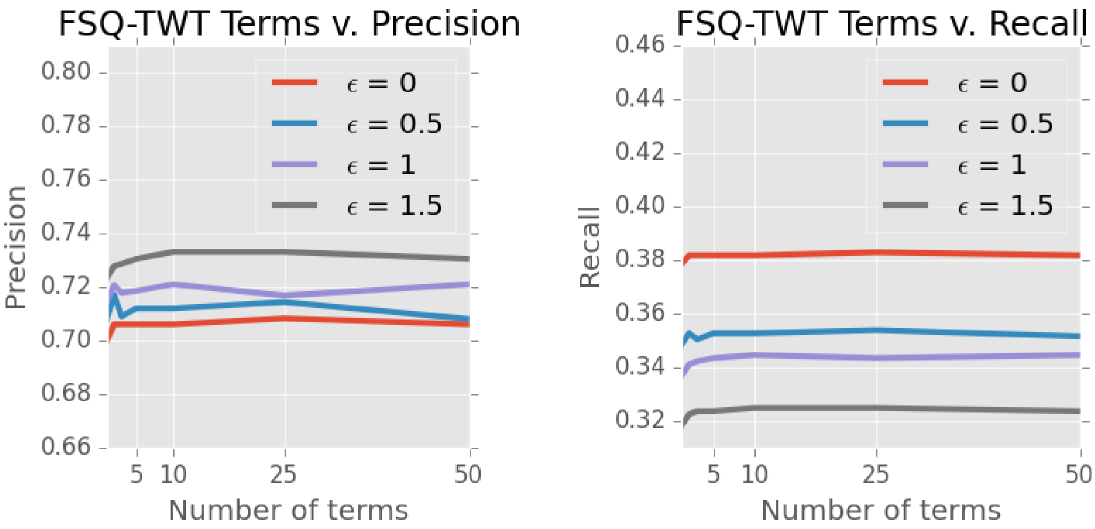


Figure 2.7: Precision and recall for the FSQ-TWT datasets for different values of the eccentricity and varying numbers of terms of the infinite sum.

privacy implications through partially or inaccurately filled profile information (such as entering a fake name, age, etc.), or using the privacy settings to "lock down" access. However, such methods are of limited use, because commonly collected fields (such as location) that are integral to the service provided may *in themselves* be sufficient to link this account with other accounts of the same user.

In this paper, we present a new approach to characterize when and how such linking is possible. We theoretically justify our algorithm and empirically validate it on real datasets. The results we present, most of them shown for the first time in a cross-domain setting, demonstrate that simple conditions may be sufficient for correct reconciliation and highlight the sensitivity of location data. Several avenues for further research are suggested by these results: Our model assumes very simple behavior by users, modeling them as generating location records independently, and is already quite effective. Can one further exploit patterns inherent to human mobility, such as sleep schedule, commute patterns, working days, and other time dependencies? Is location special, or are there other universal characteristics that are equally meaningful?

Chapter 3

# *Inferring Demographics from Location Data*

In the previous chapter, we discussed the re-identification risk of location data. The lack of anonymity in location data has been widely reported, but the *discriminative* power of mobility has received much less attention. In this chapter, we fill this void with an open and reproducible method. We explore how the growing number of geotagged footprints left behind by social network users in photosharing services can give rise to inferring demographic information from mobility patterns. Chiefly among those, we provide the first detailed analysis of *ethnic* mobility patterns in two metropolitan areas. This analysis allows us to examine questions pertaining to spatial segregation and the extent to which ethnicity can be inferred using *only* location data. Our results reveal that even a few location records at a coarse grain can be sufficient for simple algorithms to draw an accurate inference. Our method generalizes to other features, such as gender, offering for the first time a general approach to evaluate discriminative risks associated with location-enabled personalization. The work in this chapter was presented at the Conference on Social Networks in 2015 [101], with work contributed from Sebastian Zimmeck, Coralie Phanord, Augustin Chaintreau, and Steve Bellovin.

## 3.1 Motivation and Summary of Results

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 [118]." 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 [84, 41].

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 [98] and displaying more relevant advertising [69] 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 chapter 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 §3.2, 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. (§3.3).
- 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. (§3.4).

- 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. (§3.5).

## 3.2 Background

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 [18, 36, 91]. In particular, we aggregate location data into mobility patterns and compare our patterns to those published in earlier studies [6, 51, 49] 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 [17] and community well-being [62] measured at city and neighborhood levels. While some studies already demonstrated that mobility traces can uniquely identify individuals [84, 110], 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 [24], activity type prediction [67, 71], and identification of location types [50].

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 [15, 79], or Tweet content [98, 99]. Those studies demonstrated an underrepresentation of females and minorities online [79]; a finding which we extend and confirm using photosharing services. Mobility data from mobile phones were used to predict personality traits [82], age [11], and gender [109], 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 [2], 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 [22]. Probably, the closest work is [137], 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 [117] (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., [84, 62]). 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.

## 3.3 Methodology and Application

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 3.1.

**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 [50, 51, 49] 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' ethnicities 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 [73] and that 46.6% of photos are either selfies or show the user posing with other friends [47] 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 [93], 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 3.2. 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 $\alpha$ [58]. 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 [74]. Figure 3.2 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 §3.4 we used data from the Census [117] 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 by ethnicity- and gender-specific Internet [30, 72] and Instagram [25] usage numbers. As explained in §3.4 we also took into account that Caucasian Hispanics are often perceived as Caucasian alone [77]. 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 (§3.5) since those features could provide an estimate of the types of places a user would visit.

## 3.4 Mobility-Demographics

We now present a mobility pattern analysis for various population levels. Our dataset reveals mobility trends similar to those of CDRs (§3.4) and generally represents the adjusted Census population well (§3.4). 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 (§3.4), and we are also able to detect segregation among ethnic groups (§3.4).

### Mobility Patterns

In order to compare the mobility patterns of our dataset to those in the CDR dataset of [51, 49] 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 3.1 shows the distribution of the data in our subsets compared to those in the CDR dataset [51]. 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—

| | *Spring* | | *Winter* | |
|---|---|---|---|---|
| *Statistic* | LA | NY | LA | NY |
| Total Checkins | 135,503 | 109,506 | 118,446 | 98,286 |
| (Total CDRs) | (74M) | (62M) | (247M) | (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 3.1: Statistics of our LA and NY subsets compared to the CDR dataset in [51] (where available, in parentheses). Our calculations do not consider any day where a user had no checkins.

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 [49]. 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 [49]. Figure 3.3 shows a subset of our results. Our ranges are generally smaller than those reported by [51, 49]. 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.

## Demographic Patterns

As our LA and NY subsets are annotated with ethnicity and gender labels (§3.3) 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 [117] 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%) [30] and Instagram (16% vs. 10%) [25] 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 [72]. 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 [79]. We found this phenomenon difficult to assess, specifically, as ethnicity is not significant for setting a profile private [65], activity levels (posting pictures, etc.) are not lower for Hispanics [113], 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 [77]. 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 3.4, 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.

### 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 3.5 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 3.3, 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 3.5, 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 [4] 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 §3.3, we calculate the distance between the home and each checkin for the different ethnic groups and genders. Figure 3.6 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 [55] and travel longer distances to work when they are employed full-time [61]. 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 3.7 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.

## 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 [75]. 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) [100]. 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 [75]. 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 [75]) and the entropy index measures the dimension of evenness (the extent to which minority group members are over- or underrepresented in an area [75]). The interaction index, $B$, can be understood as the probability of a minority group member interacting with a majority group member and is defined [124] by

$$B_{kl} = \sum (\frac{n_{ik}}{N_k})(\frac{n_{il}}{n_i}), \tag{3.1}$$

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 [22] 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 [124], $H$, as

$$H = \frac{H^* - \bar{H}}{H^*}, \tag{3.2}$$

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

$$H^* = -\sum_{k=1}^{K} P_k ln(P_k), \tag{3.3}$$

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

$$\bar{H} = -\sum_{i=1}^{I} \frac{n_i}{N} \sum_{k=1}^{K} P_{ik} ln(P_{ik}), \tag{3.4}$$

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 3.4, 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 (§3.3) and that we adjust label and Census distributions (§3.4).

**Residential Segregation** Tables 3.2 and 3.3 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 [48]. 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

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

Table 3.2: 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 3.4).

| *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%) | **3%** |

Table 3.3: 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 3.2, the last column shows the measurement error. As further explained in Table 3.2, we did not consider NTA (LA) and state granularities (LA and NY).

the smaller number of data points. While the level of interaction seems to increase when areas become 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 3.3, we find another indicator for low segregation [48]. 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 3.4 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.

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

Table 3.4: 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.

## 3.5 Inferences from Mobility Data

We now show how location data by itself allows to infer ethnicity and gender of individual Internet users. We introduce a simple frequentist approach (§3.5), describe considerations informing our methodology (§3.5), and present the results of its application (§3.5).

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

Table 3.5: 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. Logistic Regression was the best algorithm for all problems. The baseline was obtained by predicting the class of a user based on the label distribution.

### 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 $\mu_1$ (respectively $\mu_2$) 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 independently from this user's class probability distribution. The prior probability that a user is in class $C_1$ or $C_2$ is denoted $\pi_1$ and $\pi_2$, 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 [38]. However, in practice humans show periodicity [36] or even social bias [18] 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 (§3.5). 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

$$\forall j,\ P(X_j^{(i)} = l | i \in C_1) = \mu^{(1)}(l), \tag{3.5}$$

so that

$$P(X^{(i)} = (l_1, \dots, l_n) | i \in C_1) = \mu^{(1)}(l_1) \dots \mu^{(1)}(l_n), \tag{3.6}$$

by independence, and applying Bayes' rule

$$P(i \in C_1 | X^{(i)} = (l_1, \dots, l_n)) = \frac{1}{1 + \frac{\pi_2 \mu^{(2)}(l_1) \dots \mu^{(2)}(l_n)}{\pi_1 \mu^{(1)}(l_1) \dots \mu^{(1)}(l_n)}}. \tag{3.7}$$

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

$$\pi_2 \mu^{(2)}(l_1) \dots \mu^{(2)}(l_n) < \pi_1 \mu^{(1)}(l_1) \dots \mu^{(1)}(l_n) \tag{3.8}$$

or, equivalently, if we have

$$\sum_{k=1}^{n} \ln \frac{\mu^{(1)}(l_k)}{\mu^{(2)}(l_k)} > \ln \frac{\pi_2}{\pi_1} . \tag{3.9}$$

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,

$$P(\text{error} | n \text{ checkins}) \approx_{n \to \infty} 2^{-n\mathcal{C}(\mu_1, \mu_2)}, \tag{3.10}$$

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

Based on this analysis, a simple algorithm to infer ethnicity or gender can first estimate $\mu_1, \mu_2$ and $\pi_1, \pi_2$ using the training data and then classify according to this likelihood rule.

## 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 §3.3); 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 [95] 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 [84, 132]. 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.

## Results

The results of our best-performing algorithms are displayed in Table 3.5, and a detailed comparison of accuracy as a function of granularity can be seen in Figure 3.8. 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 [98] 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 [2] achieved an F-measure of 0.81 and for gender an F-measure of 0.61. For gender [137] 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 3.8 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 3.9. 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.

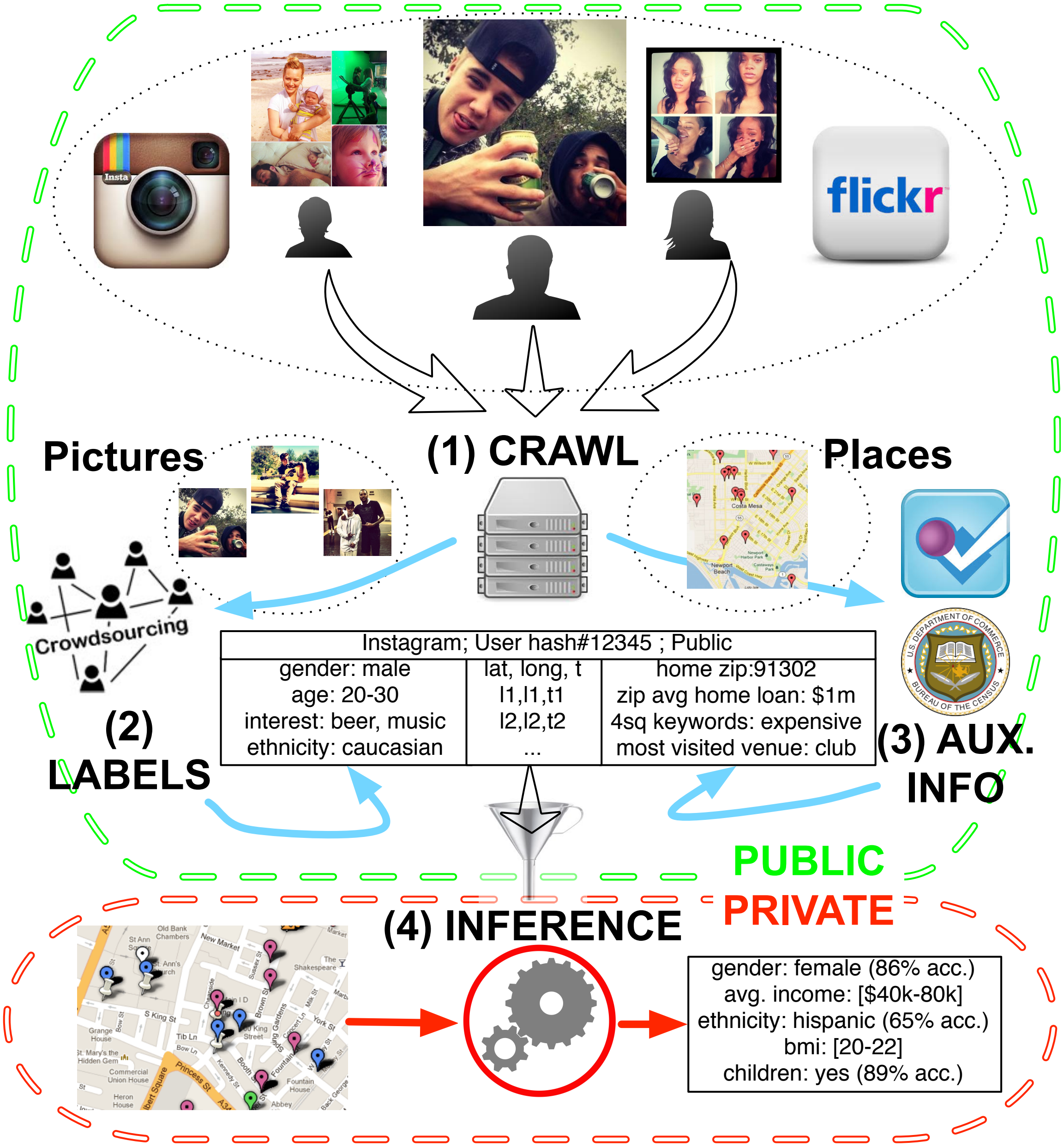
We also found that when a user is only observed in a limited set of locations, the inference accuracy increases fast with a relatively small increase in the number of locations. Moreover, it is not even required to focus on the most significant locations of a user to get good inference accuracy. Observations 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 3.10, suggest that inference for the purpose of ethnicity identification is quite robust to data sparseness and obfuscation methods.

## 3.6 Conclusion

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 §3.4, 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.

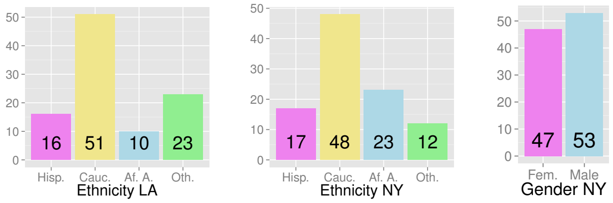
This work motivates multiple avenues of further research: First, it enables the extension of demographic mobility analysis to many researchers using shareable public datasets and reproducible results. Beyond ethnicity and gender, attributes such as age, occupation, and other lifestyle features may be 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.



| Instagram; User hash#12345 ; Public | | |
|---|---|---|
| gender: male<br>age: 20-30<br>interest: beer, music<br>ethnicity: caucasian | lat, long, t<br>l1,l1,t1<br>l2,l2,t2<br>... | home zip:91302<br>zip avg home loan: $1m<br>4sq keywords: expensive<br>most visited venue: club |

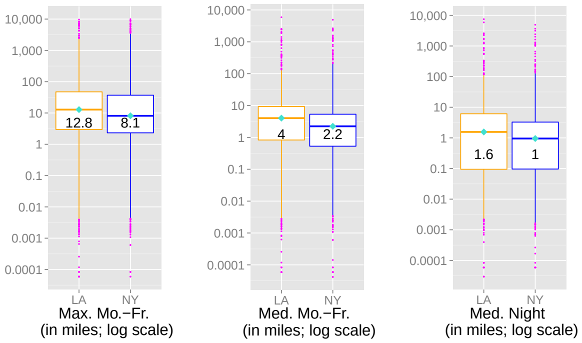
gender: female (86% acc.)
avg. income: [$40k-80k]
ethnicity: hispanic (65% acc.)
bmi: [20-22]
children: yes (89% acc.)

Figure 3.1: 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.



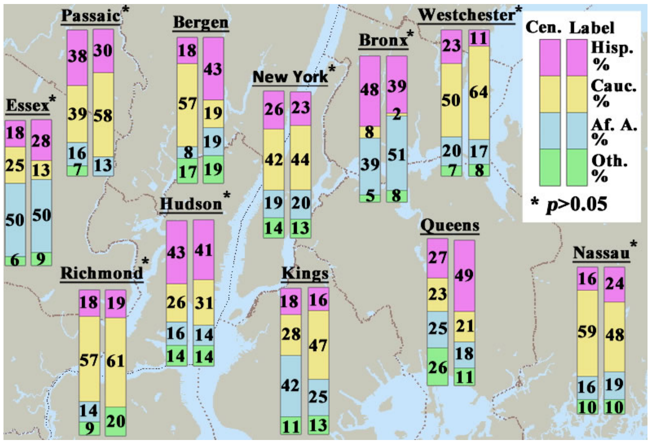
| | *Ethnicity LA* | *Ethnicity NY* | *Gender NY* |
|---|---:|---:|---:|
| *Users (n)* | 427 | 588 | 241 |
| K.'s $\alpha$ Multi. | **0.74** | **0.68** | - |
| K.'s $\alpha$ Binary | **0.78** | **0.74** | **0.85** |

Figure 3.2: Annotations for LA and NY. Top: percentages of user labels for the different categories. Bottom: absolute numbers of labeled users and annotation agreement results.



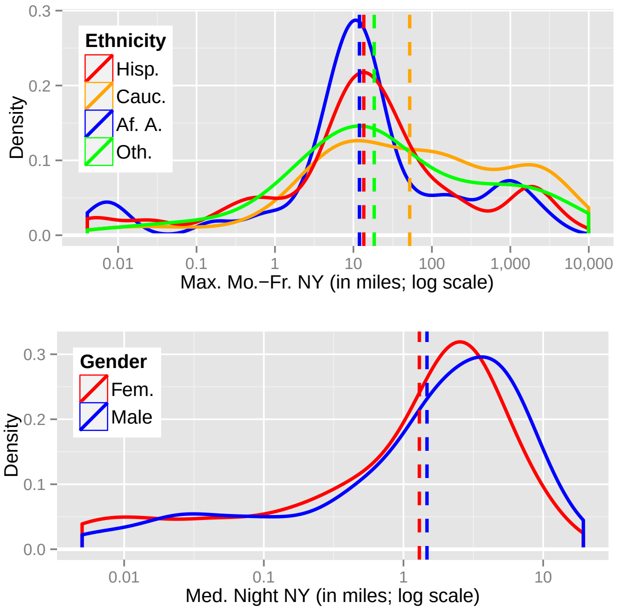
| | *Max. Mo.–Fr.* | | *Med. Mo.–Fr.* | | *Med. Night* | |
|---|---|---|---|---|---|---|
| % | LA | NY | LA | NY | LA | NY |
| 98 | 2,471.7 | 3,625.6 | 133 | 209.9 | 117.4 | 129.9 |
| | (2,467) | (2,455) | (32) | (29) | (23.1) | (19.4) |
| 75 | 47.5 | 37 | 9.3 | 5.3 | 6.1 | 3.3 |
| | (130) | (111) |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

Figure 3.3: 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 [49]. 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 [51]. Previous results [51, 49] 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<br>(0%) | 0/1<br>(0%) | 1/1<br>(100%) | 0/1<br>(0%) | 1/1<br>(100%) |
| County | 1/2<br>(50%) | **8/11**<br>**(73%)** | 2/2<br>(100%) | 6/8<br>(75%) | 4/4<br>(100%) |
| PUMA | 12/16<br>(75%) | 11/17<br>(65%) | 2/2<br>(100%) | 5/6<br>(83%) | 1/1<br>(100%) |
| NTA | -<br>- | 9/16<br>(56%) | -<br>- | 7/7<br>(100%) | 2/2<br>(100%) |
| ZIP | 3/3<br>(100%) | 8/14<br>(57%) | 1/1<br>(100%) | 3/3<br>(100%) | -<br>- |

Figure 3.4: 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 [107] 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. Mon.-Fri. 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 | $\epsilon$ | $\epsilon$ | $\epsilon$ | $\epsilon$ | $\epsilon$ | $\epsilon$ |

Figure 3.5: 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.

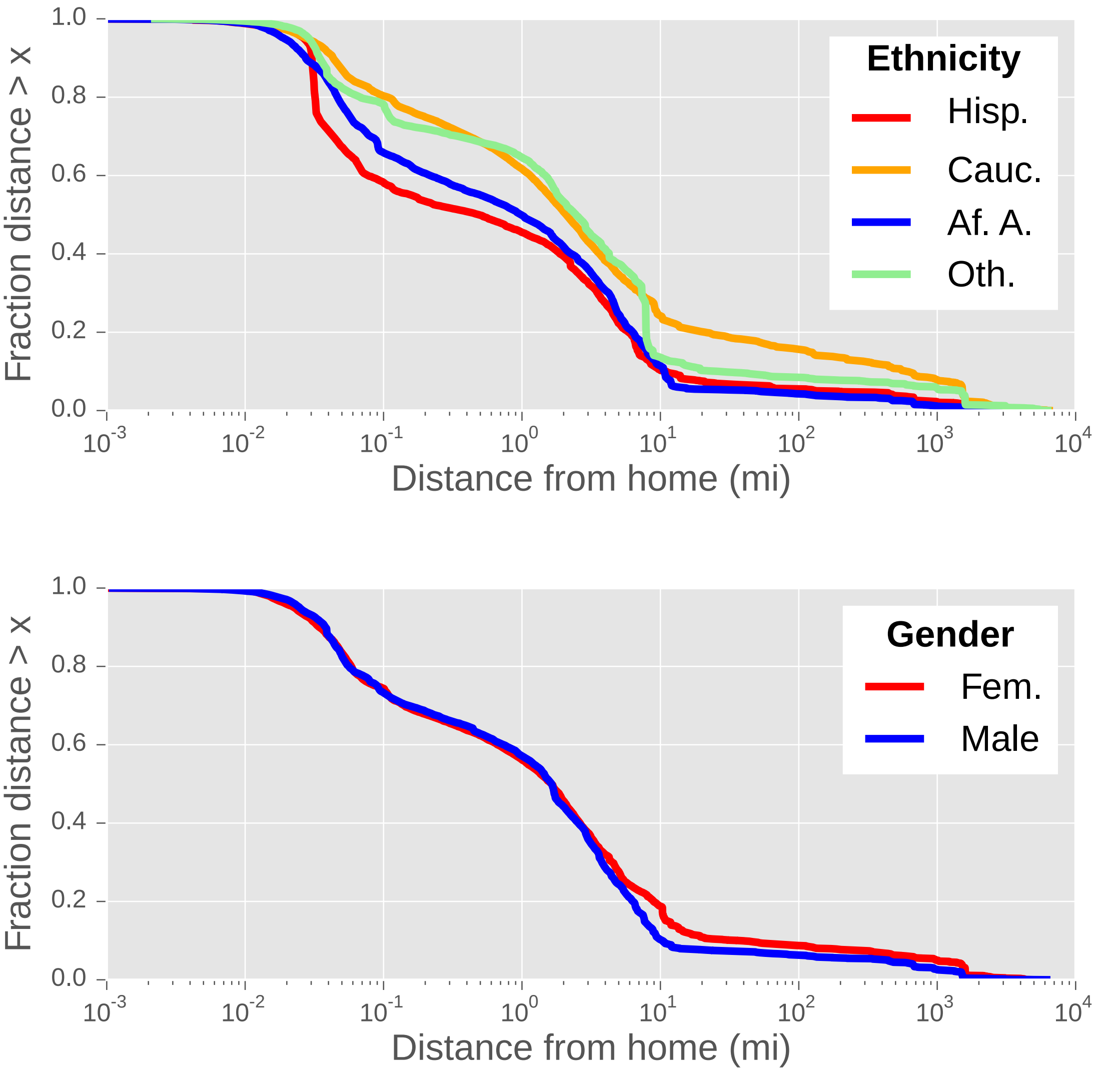


Figure 3.6: CCDFs of home ranges for NY. Top: CCDFs for different ethnic groups. Bottom: CCDFs for males and females.

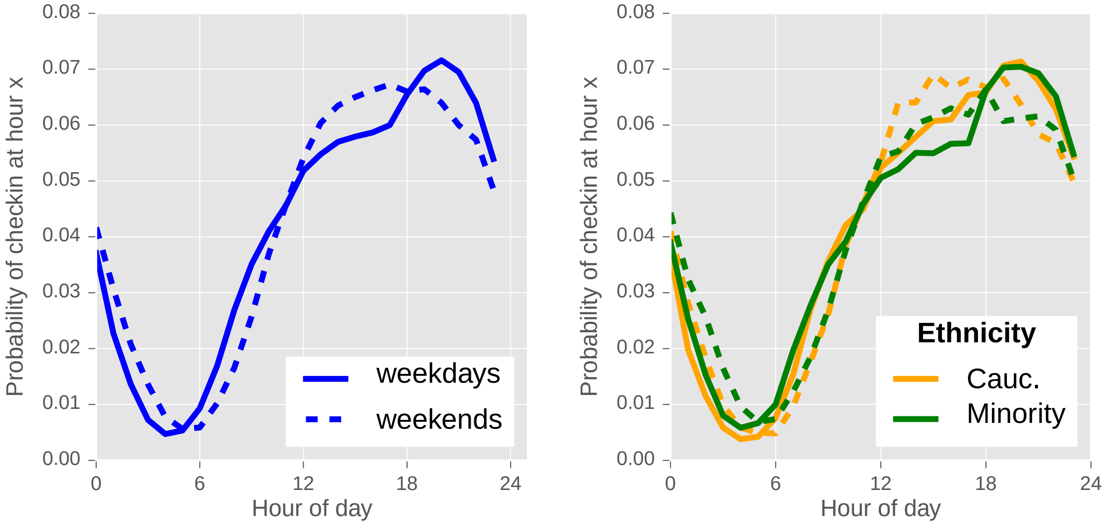


Figure 3.7: Histograms of checkin times for NY. Left: Comparison of weekends and weekdays for all user groups. Right: Comparison of Caucasian and minority user groups for weekends and weekdays. Dashed lines correspond to weekends, solid lines to weekdays.

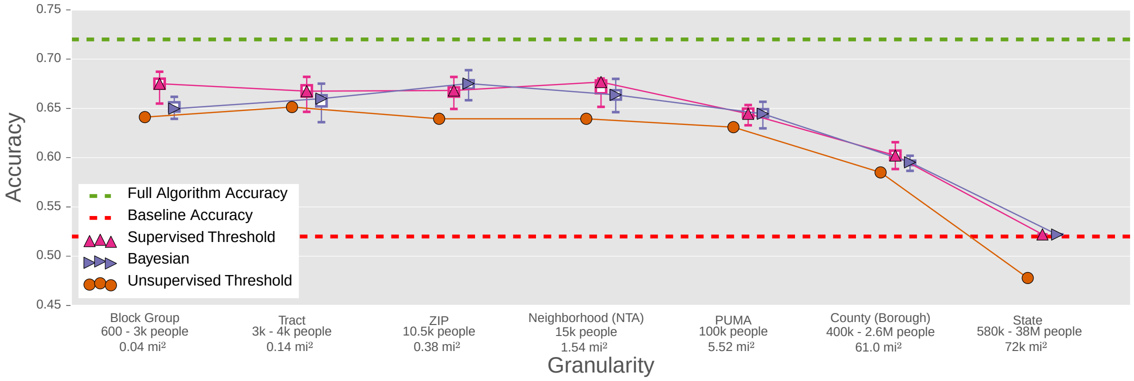


Figure 3.8: Accuracy of ethnicity prediction versus granularity for our NY population using several different inference techniques. Accuracy increases slightly at the ZIP code and neighborhood granularities and then decreases. Interestingly, the Bayesian algorithm, which uses only counts of visits to locations, performs comparably to the Supervised Threshold algorithm, which uses data on the ethnicity of visited locations.

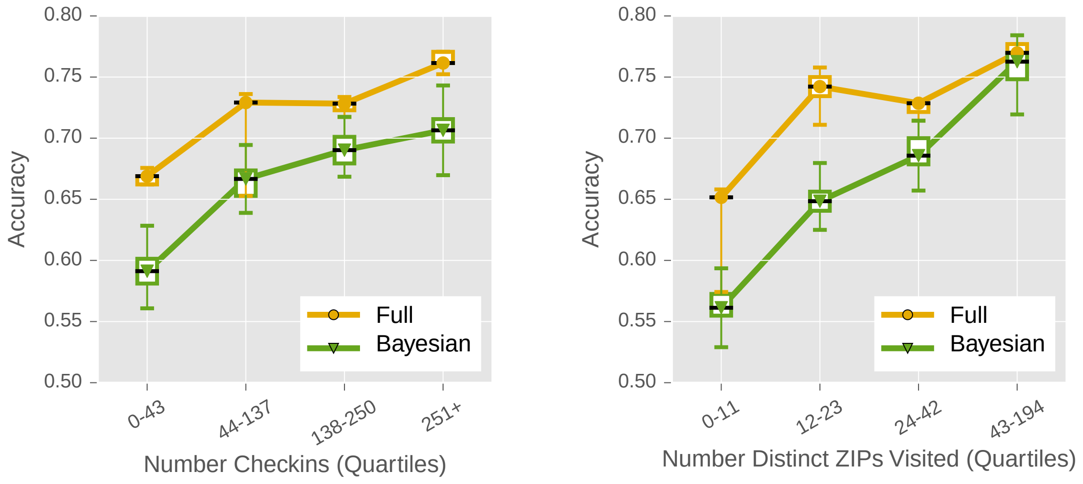


Figure 3.9: Checkin user activity. Left: accuracy as a function of total number of checkins at ZIP code locations. Right: accuracy as a function of number of checkins at distinct ZIP code locations.

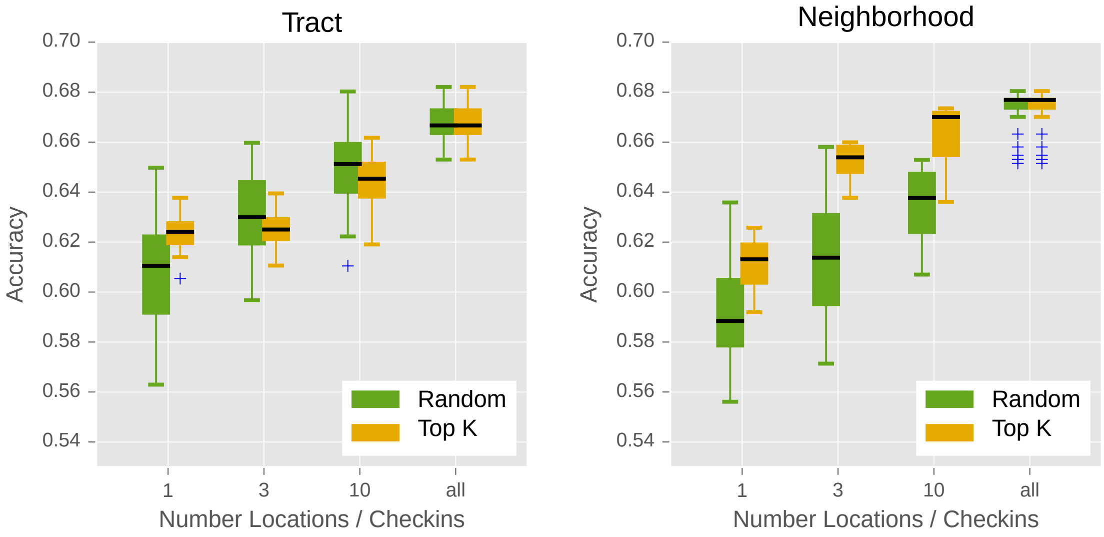


Figure 3.10: Accuracy of predicting a user's ethnicity from a small number of locations chosen either as most frequently visited locations or randomly. The algorithm used is the Supervised Threshold algorithm. Left: tract granularity. Right: neighborhood granularity.

# Chapter 4

# *Transparency in Location-Data Systems*

In this chapter, we present FindYou, a web-based application that gives users the ability to perform a location data privacy audit. FindYou lets users import and visualize the location data collected by popular web services in order to understand what these companies know or can easily infer about them. Additionally, FindYou gives users the option to donate their data to the scientific community, creating new mobile datasets linked to user properties that will be open to use by academic institutions.

What is the purpose behind FindYou? In the previous chapter, we described a method to improve user privacy while still allowing the monetization of user location data. The solution was implemented at the system level in order to protect the privacy of individuals. However, system level implementations require data aggregators to change their behaviors, and here we focus on a tool that individuals can use. Additionally, as we showed in Chapter 2, the large scale collection and use of location data also creates concerns about bias at a group level as opposed to only attacks on individuals. As concerns of algorithmic bias have grown, the research community has focused on a method to both understand what data aggregators are doing and explain these methodologies to average users. The work in this chapter is in that vein of accountability and transparency research.

## 4.1 Motivation and Summary of Results

As stated throughout this thesis, the overall economic model of network-related services is that users receive free services and software from online providers. In return, the providers obtain revenue by displaying ads to users. Typically, providers only are paid when ads are clicked, or for showing ads to users within specific demographic groups that advertisers wish to target. Thus, providers have a strong incentive to deeply understand users, in order to show them the best ads or to prove to advertisers what demographic groups are seeing ads. This can create a problem when users are not fully informed about what data is being collected about them, what this data is being used for, or with whom this data is being shared. This issue has been exacerbated by the rise of smartphones– mobile technology has both made digital interactions constantly available, while also functioning as remote sensors, collecting detailed information on users' real-world movements and behaviors. One important subset of this data is location data, which details where a user was at a specific time. Users are often incentivized to share their location data, for example, with an online service to find recommendations for nearby businesses, most often with their cell phone, but also on other devices through IP-geolocation or different methods.

Online service providers can use the data for personalization, such as guessing what language the user will want to see or tailoring content to specific users. However, this data can also be used in ways that users may not be comfortable with. For example, location data can be used to infer a user's race, gender, or uniquely identify them from anonymous data sets [137, 103, 132]. Journalists have even found evidence that location data has been used in price discrimination. In one example, a newspaper found evidence to suggest that the a company was changing the prices of products purchased online based on the inferred distance of a customer to a competing store [121]. In another, Mac users were shown more expensive hotels on a travel website [76].

In response to some of the problems with the overall economic state of the web, the community has created tools to detect and measure online personalization and ad-targeting [63, 127]. These tools, though very useful, are often not designed to inform non-technical users of the problems inherent in personalization.

In this chapter, we focus privacy understanding tools on location data to create a personal location data auditing tool. This tool allows users to (1) enter or import personal location data gathered by three popular online services, (2) visualize this data, (3) view the demographics of their visited location in terms of race, income, age, and family make-up, and finally (4) receive a prediction of their demographics based on this data. We design this tool with the goal that it will be approachable and informative for all users, especially those without deep technical knowledge. Another key part of this tool is to improve future research on demographics and mobility by allowing users to donate their data.

In the following sections, I will describe some background, the overall goals of the project, and the principles focused on while designing it. The work in this chapter was presented as a demo at the World Wide Web conference in 2016 [105] and was completed with Danny Echickson, Stephanie Huang, and Augustin Chaintreau.

## 4.2 Background

This work in this chapter lies at the intersection of two areas: location privacy and computational “auditing” tools.

Location privacy is a rich field that explores privacy problems created in the use of user location data and potential solutions. Previous works have shown that location data can be used to infer sensitive traits of individuals [137, 103]. Other works have explored how users understand and value their location privacy [112]. In constrast to these works, we do not utilize user data as an object of study, or seek to understand user perceptions of location privacy. Rather, we wish to inform users about their location data and potential privacy hazards by providing the user with a visualization of their already collected data, along with what this data might suggest to a third party.

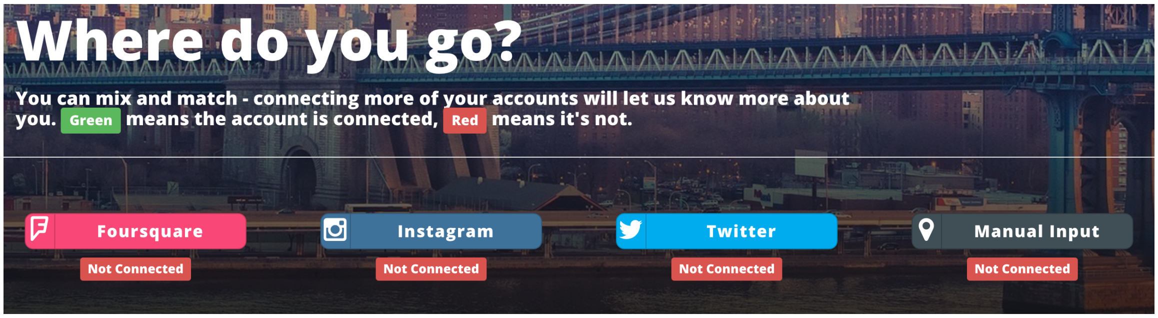
Another related collection of work is that on systems for understanding how online personalization takes place. These works have attempted to measure personalization [44, 127], price discrimination [78, 121, 76], and ad targeting [70, 63, 125]. We are closely related in that our work is concerned with these issues. However, rather than attempt to detect these problems, FindYou functions as a tool to make users aware of the existence of these issues.

There are multiple sources for capturing and visualizing your data online [37, 86]. Our work goes beyond visualization by also showing predictions informing users of what their data could be used for. Additionally, there are other projects where users can donate their data to science [34]. Our project focuses on a specific subset of this larger goal, but offers a type of data that is not publicly widely available.

## 4.3 Description

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.

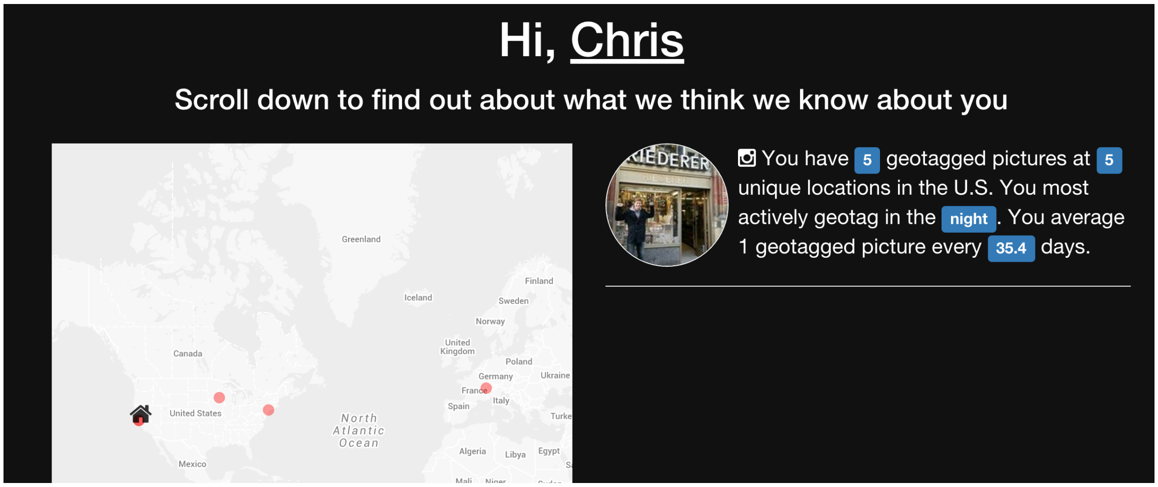
We will begin with a summary of a typical use of FindYou, and proceed to explain each component in more detail, along with the decision-making that influenced the design.



Where do you go?

You can mix and match - connecting more of your accounts will let us know more about you. Green means the account is connected, Red means it's not.

Figure 4.1: The user is presented with four different ways of connecting his or her location data to the app.



Hi, Chris

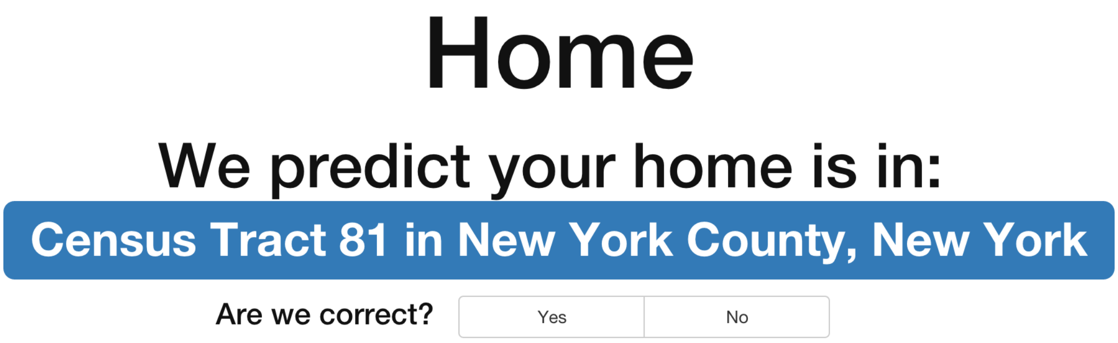
Scroll down to find out about what we think we know about you

You have 5 geotagged pictures at 5 unique locations in the U.S. You most actively geotag in the night. You average 1 geotagged picture every 35.4 days.

Figure 4.2: After connecting their data, the user sees an overview of their locations and imported data.

## 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.



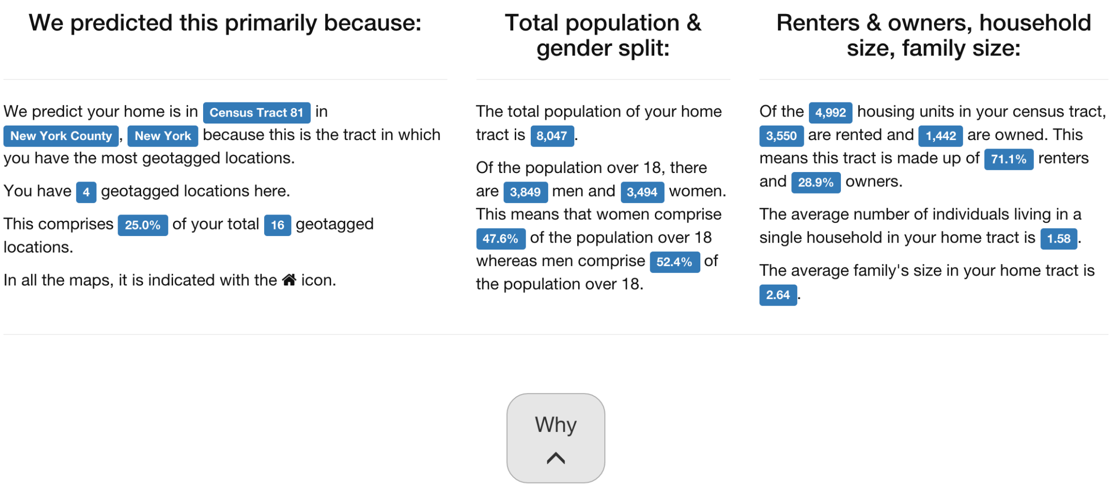
Home

We predict your home is in:

Census Tract 81 in New York County, New York

Are we correct?

Figure 4.3: We show a specific guess for the user's home location.



We predicted this primarily because:

We predict your home is in Census Tract 81 in New York County, New York because this is the tract in which you have the most geotagged locations.

You have 4 geotagged locations here.

This comprises 25.0% of your total 16 geotagged locations.

In all the maps, it is indicated with the 🏠 icon.

Total population & gender split:

The total population of your home tract is 8,047.

Of the population over 18, there are 3,849 men and 3,494 women. This means that women comprise 47.6% of the population over 18 whereas men comprise 52.4% of the population over 18.

Renters & owners, household size, family size:

Of the 4,992 housing units in your census tract, 3,550 are rented and 1,442 are owned. This means this tract is made up of 71.1% renters and 28.9% owners.

The average number of individuals living in a single household in your home tract is 1.58.

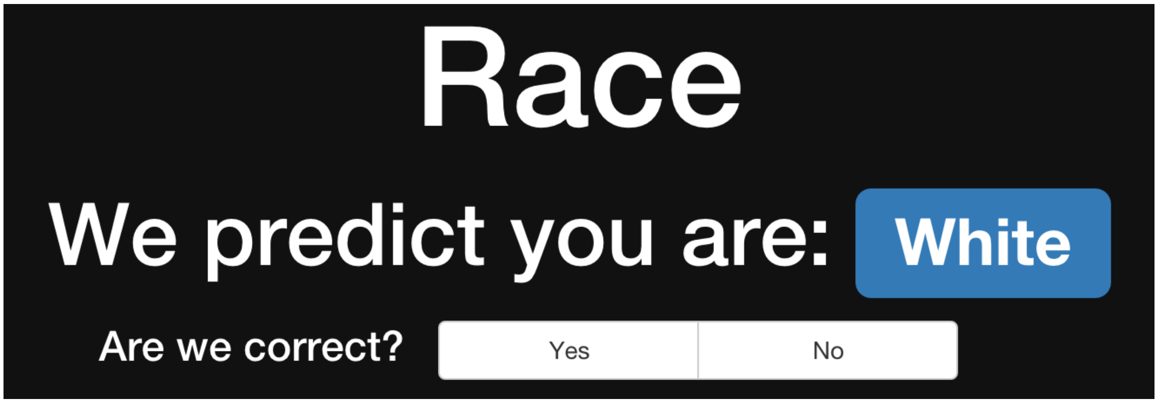
The average family's size in your home tract is 2.64.

Figure 4.4: For all predictions, we show additional details about how we made this guess.

## Design Decisions

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

**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.



Race

We predict you are: White

Are we correct?

Figure 4.5: The site predicts several demographic attributes, one of which is race. The user has the option to tell us if we are correct.

**Instagram** is a photo-sharing application owned by Facebook with 400 million monthly active users. Instagram is notable for it being primarily targeted at mobile phones; currently users cannot upload photos from a desktop or laptop computer. The mobile focus makes it is easy for users to "tag" photos with locations using their phone's GPS device. Although many users do tag their photos with location data, unlike Foursquare, it is not necessary to post a location in order to use the app. Due to the fact that many users do tag their photos with locations, it is the second-most "location-centric" of our three services.

**Twitter** is a microblogging service where users post 140 character texts called "tweets". Twitter has approximately 320 million users. Through its smartphone interface, Twitter users can tag tweets with locations. Many users connect their Twitter account to other web services, such as Foursquare and Instagram, among others, which may also contain location data. The primary focus of most tweets is not about where a user currently is. Therefore, Twitter is the least location-centric.

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

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

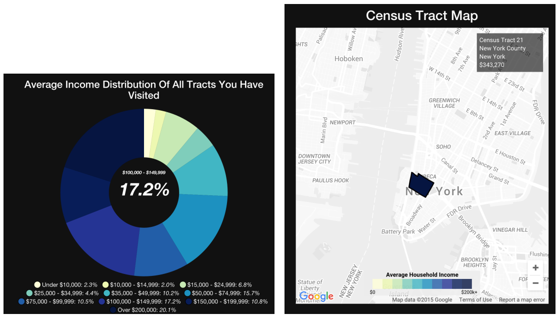
*How did you choose to represent locations?* There are many different ways to represent locations, such as latitude longitudes, venues, cities, or points of interest. Throughout this chapter and the site, we use a United States Census tract as an "atomic" location. The United States Census partitions 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.

## 4.4 Future Work

Our most important future task is to obtain widespread usage and determine the most useful features of the site. FindYou is currently public and live. By showing it to more users, we hope that we can obtain valuable feedback and to rapidly iterate to present an engaging and informative perspective on the gathering of location data. One possibility is to run randomized controlled trials with FindYou and assessing its effect on attitudes or awareness of privacy issues.



(a) (b)

Figure 4.6: (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.

Multiple improvements can be made to the site. We would like to offer more support in diverse geographic regions outside of the United States. Additionally, we could expand to other popular services or to more advanced forms of data uploading such as GeoJSON or text files of latitude-longitude pairs. Another possible improvement would be to expand the number of demographic traits on which we classify, or to use more advanced classifiers.

We look forward to sharing any data that we obtain with the research community in a way that both protects the data of donating individuals as well as making it easy for members of the research community to make new discoveries.

## 4.5 Conclusion

We have presented the motivation, design, and implementation of FindYou, a personal location privacy auditing tool. FindYou displays to the user their location data that has been collected by popular online services. Additionally, FindYou informs the user on the demographic make-up of the areas that they have visited, and shows how this data can be used to infer traits about the user. In addition to these web services, FindYou allows users to manually edit their location data to see the impact of adding and removing locations on these predicted traits. FindYou allows users to donate their data, with the hope that eventually the research community will have a useful set of user location histories tagged with demographic information. The site is currently live at `https://find-you.heroku.com`.

Chapter 5

# *User Choice in Location Disclosure*

In the first two chapters, we showed that location data is difficult to truly anonymize and can additionally reveal the demographics of users. What is a privacy-conscious user to do? In this chapter, we begin to present solutions to this conundrum.

## 5.1 Motivation and Summary of Results

As discussed in the introduction, many of the privacy concerns around location information are rooted in how the mobile application ecosystem works. Most mobile services and applications are free and operate by collecting various types of personal information about the user (browsing activity, location etc.) and monetizing this information through targeted ads [64]. Many web services and applications access location information even when such information is not needed, and may share it with multiple third parties, leading to privacy concerns [28, 122] and attracting the attention of regulators [31, 1, 118]. Location-based adveriting, however, is often quite effective: location-based targeting can garner four times as much revenue per impression compared to ads without location data[1], and even brick-and-mortar stores are interested in location data, with retailers using cell phones' WiFi signals to learn about where customers spent time in their stores[2]. This is why, when privacy advocates request stricter rules to be enforced on information collection, they typically are opposed by companies providing these services. Apps and service providers claim

[1] `http://bit.ly/vXWdsw`

[2] `http://nyti.ms/15vLRva`

that the "cost" of a privacy bill threaten the web's general economy and, ultimately, hurts customers. An ideal privacy solution, therefore would provide adequate privacy protection to the user and at the same time enable the service providers to collect and monetize data.

In this chapter, I describe a system that gives users control over their information, does not degrade the data given to aggregators, and preserves revenues. Recognizing that the first challenge is to express locations in a way that is meaningful for advertisers and users, I propose a *keyword* based design. Keywords characterize locations, let the users inform the system about their sensitivity to disclosure, and build information directly usable by an advertiser's targeting campaign. This chapter has three contributions: the design of a market of location information and discussion of its robustness; an analysis of the economic consequences of the system using data from ad-networks, geo-located services, and cell networks; and a small scale experiment of the system to collect preliminary results on the behavior of real users if location information markets were deployed.

The objective of this chapter is to lay the groundwork for a comprehensive and deployable solution to location privacy. In contrast to previous works, we aim at reconciling the control users exert over their data with its commercial value. This raises three main challenges. The solution should be *incrementally deployable*: it must easily integrate with current devices and practice, while giving all parties an incentive to participate. The solution should be *robust* against threats from its participants. Advertisers wishing to access data without compensating users, or access more than the users specify, should be stopped. Users should not be able to significantly benefit from seeking unfair compensation. The solution should be *easy to use*: users and advertisers have to express their needs in intuitive terms.

Our solution is based on selective disclosure; users decide which 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 (more in Sec. 5.2). In order to drive the adoption of the solution, I propose to include 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.

The main contributions are:

- 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).
- An analysis of how such a system can offer different levels of protection against various threats, including free riding, inference attack using auxiliary information, and user misconduct (Sec. 5.3).
- An evaluation of a deployment within the economy of mobile advertising. We use data gathered from cell phone users, geo-located services, ad-networks, and a simple revenue model. We found multiple privacy-value trade-off that benefit users and advertisers. We find that if information is removed about most privacy sensitive locations, revenue drops by around 20% (Sec. 5.4)
- 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. 5.5).

Much of this chapter appeared as a paper at the Workshop on Privacy in the Electronic Society in 2013 [102]. The work was conducted with Augustin Chaintreau, Vijay Erramilli, and Jacob Cahan.

## 5.2 Overview

In this section, we discuss our assumptions, provide a description of the design, give a simple example to explain how our solution works, enumerate the advantages of the system, and describe the data we used to analyze the solution.

### Requirements and Assumptions

To meet our requirements, we create a solution based around *selective disclosure*; users disclose location information that they are willing to release, and this information can be monetized by ad-networks and third party aggregators by way of online ads. The control, therefore is with the user. Any privacy solution based around selective disclosure (Koi [42] etc.) needs to address *how* the information is released, under *which conditions* the information is released and to *whom*.

In order to answer *how* we release location information, we design our solution around keywords associated with locations; the decision to release is based on keywords associated with locations, while the information that is actually released is the location. A simple 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 latitude longitude pair (lat-long) of each particular venues. This association has several advantages: (i) Keywords let us deal with the problem of location privacy at a higher abstraction than coordinates or even location descriptors (as Koi [42] does). (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.

In order to answer *under which* conditions we release, users opt-in and disclose location information of their own choice[3], and they get compensated *economically* for this release by aggregators and ad-networks. This agreement is facilitated by a trusted third party who provides access to the user at *only* those locations the user releases, upon payment. We use economic compensation for multiple reasons. In general, concerns around privacy alone have not helped in large scale adoption of privacy solutions; past research has shown that most users fall prey to cognitive biases while thinking about privacy solutions [10]. We hope economic incentives can nudge more users towards adoption. Introducing an economic dimension to the disclosure problem also addresses the issue of to *whom* the information will be released to – parties that can pay. In addition, the trusted third party in the middle can vet the parties.

Before describing the design, we first describe our **assumptions**. We consider our adversary to be an honest-but-curious advertiser. This means our adversary participates in the system honestly but may try to exploit the information that is gathered. With this in mind, we provide safeguards against inference and linkage attacks.

We assume that once an entity enters into an agreement with the user, it is generally compliant. Given the amount of press on privacy related issues, we believe that the PR backlash in the case of a serious privacy violation will make such violations undesirable. We assume that location information can be tracked and gathered continuously, as this is the worst case. In general this is not feasible as energy concerns around GPS usage will forbid this [68]. We note that the architecture presented next is oblivious to a background service model (passive, potentially continuous tracking)

[3] Users are comfortable disclosing location information under certain circumstances [54]

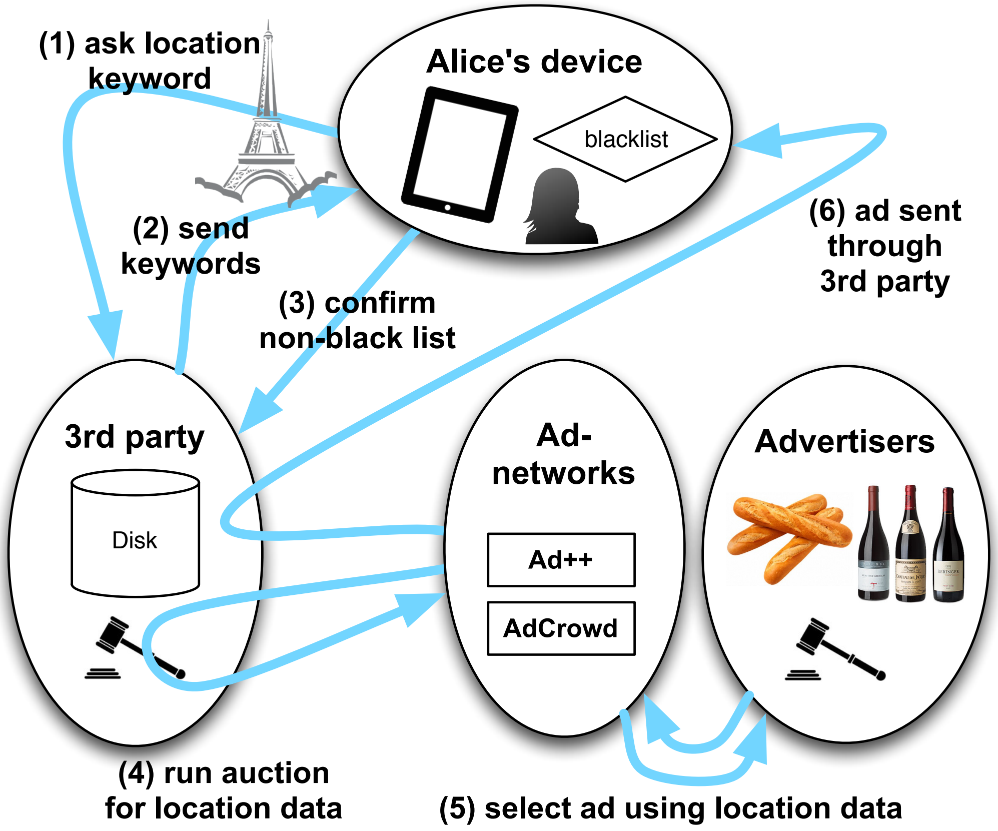

(1) ask location keyword
Alice's device
blacklist
(2) send keywords
(3) confirm non-black list
(6) ad sent through 3rd party
3rd party
Disk
Ad-networks
Ad++
AdCrowd
Advertisers
(4) run auction for location data
(5) select ad using location data

Figure 5.1: Solution overview

or a check-in model. We also assume that modern mobile OSes truly implement users' privacy preferences. If a user decides to not share location information with a certain application, then this request would be enforced.

## Design and Example

The architecture consists of the following components: (i) a blocking module in the network that blocks access to various parties, (ii) a blacklist module that contains a list of sensitive keywords and maps these keywords to physical locations – these are locations that will not be revealed, (iii) a market that puts up location information, for locations visited by the user that are not in the blacklist, (iv) an module that grants *access* to the user for parties that pay, after having come to the market and used the released location information to valuate the user. All these modules are stored in the network; *no* changes are required on the device.

The high-level diagram is shown in Fig. 5.1. We describe the process by using a simple example. User Alice is willing to share certain locations, and would like to hide her presence at other locations, a typical occurrence [54]. Alice would like to buy some bread to go with dinner, shop for wine and then head to the Libertarian party headquarters in her town to volunteer for the upcoming elections. She would like to conceal the fact that she is an active volunteer as it would disclose her political leanings. Alice, would therefore put 'Libertarian, Politics' as some keywords in her blacklist. We describe in Sec. 5.5 how the blacklist formation can be simplified through nested menus and re-ordering. This blacklist will be stored on a server at a third party location. We assume the third party is trusted and leave lowering this requirement to future work.

As Alice walks to her locations, all 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. [59] (see Sec. 5.5 for more on implementation). At every location, a check is made against the blacklist to verify if Alice is comfortable releasing this information. In order to perform this check, we need to translate the keywords to locations, described in Sec. 5.2. Once a location passes the check, as in this case, 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 ($UID_{Alice}$, ($\mathrm{lat}_1$, $\mathrm{long}_1$), Bakery). As she walks to the wine shop, the information on the market will be ($UID_{Alice}$, ($\mathrm{lat}_2$, $\mathrm{long}_2$), Bakery, Wine Shop), as the wine shop also passes the blacklist test. Ad-networks can pay to *access* Alice based on these two locations released (described in Sec. 5.4). 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, that is conveyed to the ad-networks. Alice can receive a targeted ad (via an app or via SMS) for a particular wine selection.

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.

## Mapping Locations to Keywords

Using a mapping of locations to keywords has a variety of challenges and advantages, which we now discuss. Locations may be defined in two ways which are both compatible with our system. It may denote a point of interest where users "check in" (as in services like Foursquare) or alternatively it may represent a certain geographical area (defined using lat-longs or the coverage of a given cell tower).

Creating a mapping of locations to keywords is not necessarily easy *per se*, but one can reuse online services already providing such a mapping, such as Yelp, Google Places, and Foursquare. A "folksonomy" approach could also be used where users augment the map over time, and even receive incentive. In this case, to encourage tagging of privacy-sensitive locations, the system can allow anonymous tagging. Usability is also a challenge, and care must be taken to keep the number of keywords manageable and design the blacklist's user interface to be easy to use (see our UI in Sec. 5.5).

**Multiple benefits to the user** come from mapping locations to keywords. 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.

**This mapping allow advertisers to make sense of locations**, as today's ad-networks already offer keywords to use for context (see Appendix B). 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.

**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.

## 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. Regarding adversaries who can infer the identity of the user or learn about locations on the blacklist, we deal with this form of attack in Sec. 5.3.

**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.

**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

| Data set | # Users | # Checkins | # Locations | Duration |
|---|---|---|---|---|
| Foursquare | 40,578 | 1,377,181 | 460,663 | March-Aug 2011 |
| CDR | ∼ 2 mil | ∼ 800 mil | ∼ 7000 | 3 months 2009-10 |

they do today.

## Data-Driven Approach

To evaluate potential attacks as well as investigate economic properties of our solution, we used several large data sets. We gathered (i) mobility patterns of a populations, (ii) geo-data to associate a location with a particular keyword, and (iii) estimates of the commercial value of advertising targeted at each keyword. To the best of our knowledge, no prior work ever combined them. These data sets included:

**Location data from call description record (CDR) data** for two major western European cities (referred to as city A and city B), obtained from a large European mobile provider, for a period of three months during late 2009, early 2010. A CDR is a record that is collected by mobile providers whenever a call is made by a subscriber/user. Each record contains a user identifier, time of call, and an id of the cell tower that handled the call. The data contains over 800 million different calls placed by over 2 million users at several thousand cell towers. We focus on major cities with high density, so most cell tower ranges are small (about 100m).

**Location data from Foursquare**, obtained by crawling publicly available tweets of checkins, collected between Mar-Aug, 2011. In total, our dataset had 40,578 users, 460,663 locations, and over 1.3 million checkins. Foursquare is a location based service where users "check-in" at locations. Foursquare data compliments the CDR data well, as it gives us exact, semantic knowledge of a location as opposed to GPS coordinates that could mean a number of locations (e.g. Columbia University vs. (40.8092652, -73.9612935)). Each Foursquare location is marked with a category, which we assigned to be that location's keyword.

**Associations of locations to keywords** for the several thousand cell towers in our CDR data, obtained through the Yelp API. `Yelp.com` is an online ratings and review company. One of Yelp's API calls provides information on all the businesses within a certain radius of a lat-long point. To decide what radius to use, we first partitioned the cities with a Voronoi tessellation seeded at cell towers, as is often done to associate areas with cell towers [12]. To approximate the tessellation with a circle, we identified neighboring towers with the Delaunay triangulation, and set our querying radius to be half of the farthest neighbor. We used the categories of the businesses returned by a Yelp API call to be the keywords of that region. This approach yielded 447 distinct keywords. Note that in contrast to the Foursquare data set, each location could have several different, potentially unrelated keywords. For example, a bakery and a bar could both be associated with one location. We note here that we could have used a service different from Yelp; our method is general. Yelp provides convenient APIs and had good coverage.

**Keyword monetary values** by using the keyword's cost per click (CPC)[4] to map keywords to monetary value. We gathered an estimated CPC for each of our keywords through Google's contextual targeting tool (`adwords.google.com`).

[4] We could not directly get a location's value from an ad-network. To the best of our knowledge, major ad-networks do *not* yet allow bidding for real-time locations

## 5.3 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. Attacks may come from a few different directions– advertisers trying to gain access to or information about users without paying, malicious attackers trying to undermine the privacy of users, or users trying to unfairly obtain money from advertisers.

### Attacks on the Value of User Data

There are a variety of ways ad-networks may try to take advantage of information from the system without properly compensating the user. Our system prevents an adversary economically benefiting by doing so.

First and foremost, 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 one currently used anonymous ID to an older one[5]. Our system indeed 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 did not pay is not receiving the identity and the location of the user, but a random temporary ID. Then the ad-network, although it has a rich profile of user $u$, is not able to recognize user $u$ 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 of $u$ and its location is unknown, and the profile alone does not suffice to target.

A related issue is trajectory-based profiling. If an ad-network learns the habits of

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

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 application 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 application if it does not disrupt service [46]. An adversarial app could monitor the location market and try to associate an anonymous user profile with a particular device. Combined with a 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 [64], 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, hence it loses some value over time.

## 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 [88], we now pose the following question: "Can an adversary with the full knowledge of the location information of a significant fraction of users predict with the blacklisted locations of other users with high accuracy?" We test this on our Foursquare dataset, described in Sec. 5.2. 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 $\mathrm{Sim}(u, v)$ 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:

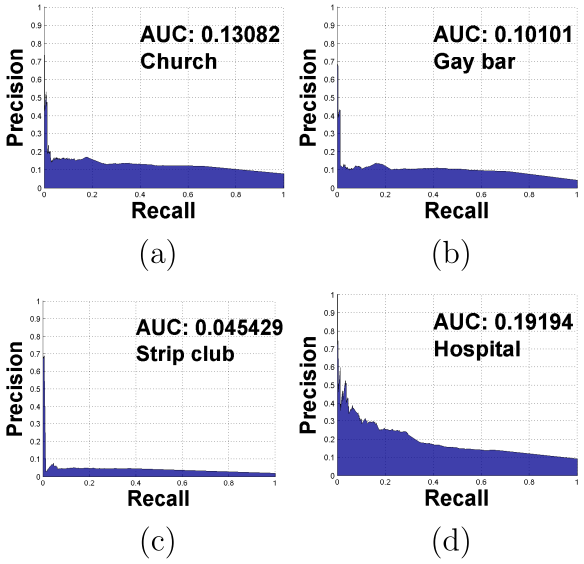
$$\mathrm{Sim}(u,v) = \sum_{l \in \mathcal{L}} \frac{1}{\mathrm{span}(l)} \mathbb{I}_{l \in L_u \cap L_v} \text{ , for } \mathrm{span}(l) = \sum_{u \in \mathcal{U}} \mathbb{I}_{\{l \in L_u\}} \text{ .}$$

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:

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

It then decide to predict 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



(a) (b)

(c) (d)

Figure 5.2: Precision-Recall curves for four sensitive keywords: (a) Church (b) Gay Bar (c) Strip Club (d) Hospital

and aggressiveness of the reconstruction technique.

We empirically study the effectiveness of this attack using 1.3 million checkins from 40,578 Foursquare users (see Sec. 5.2), in a severe case where the adversary has compromised 20% of all accounts. We vary the value of $\theta$ from 0 to 1 and plot the precision-recall of this attack for various keywords in Fig. 5.2. 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.

A few locations, at a cell level, have been shown to provide poor anonymity [85]. An interesting open question is if keywords provide better k-anonymity.

## Attacks on Advertiser Revenue

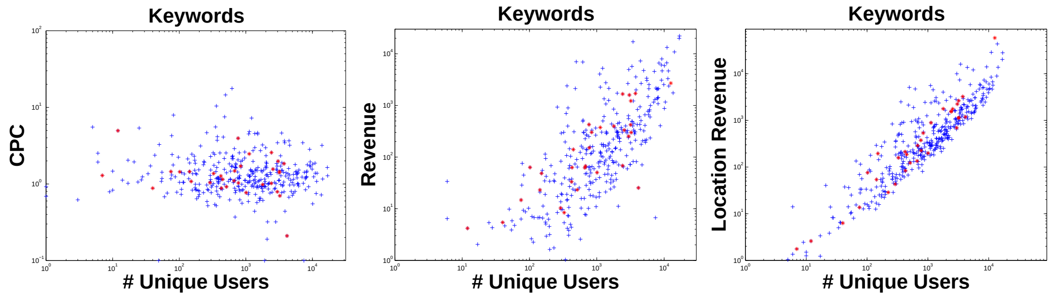
We now consider if advertisers can unfairly lose money to unscrupulous users of the system. Because users are paid when they are accessed by advertisers, they have an incentive to view or click on many ads, even when they are not interested in the displayed products, or to artificially boost their profile's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

Location fraud is actually just 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."[6] This definition applies equally well to any clicks or impressions a user creates in order to game the system. 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 location fraud. Recently, there has been a variety of research on this subject. Dave et al propose innovative methods to fingerprint click spam [23]. Haddadi suggests uses "bluff ads", ads designed to not appeal to humans and thus only be clicked by bots, to defeat click fraud [43]. Some information on the structure of Google's click fraud detection system is also available [29], [116]. Beyond the academic literature, multiple startups exist that work to estimate the rates of click fraud. These include Adometry, Visual IQ, and ClearSaleing (www.adometry.com, www.visualiq.com, www.clearsaleing.com ).

[6] www.google.com/ads/adtrafficquality/index.html

Additionally, it is easier to detect location fraud than it is to create 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. Furthermore, human beings 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 [89]. These and other 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 for a period of time or make sure the user exists through the use of a Captcha or phone call. Because of the physical constraints of location information, and because most techniques to stop click fraud can easily be applied to our system, we believe that our system is no more vulnerable to gaming than current online advertising. Although click fraud is considered an open problem by the research community, the ongoing viability of online advertising shows that our solution should likewise not be derailed by invalid traffic.

Beyond digitally generated location fraud, 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. Although this is a concern, we do not believe it to be a major issue for a variety of reasons. Traveling to a location takes significant time and effort. Such time and effort has an opportunity cost. In order to make such an attack worthwhile, the user would have to have a very valuable profile. We don’t anticipate profiles having such a high value unless the user has made multiple purchases in the past, in which case the advertiser would be compensated appropriately. Finally, the market should help deal with these attacks. We believe that valuable profiles will be distinguishable from worthless ones. The market should then be able to appropriately price them. To aide in this distinction, some reputation scheme could be added on



(a) (b) (c)

Figure 5.3: Correlation between popularity and revenue among keywords: each point represents a keyword $k$, x-axis represents the number of unique users visiting locations associated with $k$, y-axis represents: (a) $k$'s CPC, (b) the revenue generated by ads associated with $k$, (c) the revenue generated by all ads in all locations tagged with $k$. Keywords in the blacklist are represented by red stars.

top of the user's profile. For example, a user could receive a rating based on how often they respond to advertising.

Gaming of the system is certainly an issue and an area for future study. However, we believe that such concerns are no more difficult than the current click fraud situation facing online advertising. Given that online advertising is a thriving field in spite of these concerns, we feel that gaming does not pose a disastrous risk to our system.

## 5.4 Economic Analysis

The incentive to use our system is not only in mitigating various privacy leakages and attacks, but also in extracting value for advertisers and users from location information. More generally, our system aims at operating at a point where privacy and value are balanced. This is the trade off that we now wish to analyze.

To understand this trade off, we must first estimate the ad revenue. Location information today is primarily used to improve targeting, and hence the revenue, of mobile advertising. Our solution is designed to be privacy-aware but also compatible with this business model. Estimating mobile ad-revenue is a challenge in itself and it requires us to leverage our use of keywords in a model of location value grounded in real data. Note that our goal is not to estimate the absolute value of mobile advertisement. We present a simple yet rich model that captures how revenue increases *relatively* with additional information about users.

This model makes additional assumptions only for our analysis, but the system we have designed works independently of these. In particular, in our system different advertisers can have vastly different values for the same location, or at different times. Our system runs a cost-per-click auction that computes the selling price of this information (see [106]), where advertisers are incentivized to reveal their true value. In our model, however, to be closer to the way advertising currently works, we directly use the value of the winning advertisers for each keyword, as advertised by Google. This already accounts for bidders with heterogeneous values and the relative demand for each keyword. This method exploits the fact that Google that has already run the auction for us to estimate relative revenue.

## Modeling Location Value

In a cost-per-click scenario, the expected revenue of a mobile advertising opportunity (or mobile impression) is given by the bid of the advertiser that wins the auction, multiplied by the probability that the user clicks on this ad. The key decision made by an ad-network is *which* ad(s) to show. Since various factors affect this probability (user's interest in the topic, current context) additional information on a mobile impression - such as the user's previous behavior or current location - can be extremely valuable. For instance, consider the case where no information about a user is available (a situation that our model includes): the best decision that an ad-network can make is to serve a generic ad that targets the population's *common denominator*. However, this choice can be refined if some information helps improve the estimate of the chance that a user is interested in ads associated with certain keywords [128].

This permits extra revenue that our model below captures and that we claim is the *effective* value of this information.

**The revenue of a single ad** is modeled according to the principles above. Our model assumes that two effects will dominate: (i) A long term *behavioral factor* that is dependent only on the user. Studies on behavioral targeting showed that users who explicitly searched for a term in the past are more likely to react even later to an ad relevant to that term [35, 128]. Similarly, we assume mobile users who spend time in locations associated with a particular keyword implicitly display an intrinsic interest in it. In particular we define the *exposure* of a user $u$ to keyword $k$ as the fraction of time a user spend in a location relevant to that keyword, where we assume that the time spent on a location associated with multiple keywords is equally shared among these, and we denote it by $\tilde{X}_u(k)$. (ii) A short term *contextual factor* that primarily depends on location. Studies on online display advertising indicate that the quality and relevance of the page where an ad is shown matters strongly for user perception. Similarly, the ad of the winning bid associated with a keyword $k$ may not be as effective if the location where it is shown is not relevant to keyword $k$. Without loss of generality we assume that an ad out of context is only $\delta_k$ as effective as in context, where $\delta_k$ is a constant in $[0; 1]$.

Denoting by $\texttt{CPC}(k)$ the value of the winning bid for keyword $k$, and assuming that a user $u$ click ratio is dependent on its exposure $\tilde{X}_u(k)$ linearly (see Appendix B), the revenue of an ad for keyword $k$ in location $l$ for user $u$ is:

$$\mathrm{R}(u, l, k) = \begin{cases} \texttt{CPC}(k) \cdot \tilde{X}_u(k) & \text{if location } l \text{ is relevant to } k \\ \texttt{CPC}(k) \cdot \tilde{X}_u(k) \cdot \delta_k & \text{otherwise.} \end{cases} \tag{5.1}$$

**The value of location information** is how much it allows the ad-network to select $k$ to increase the revenue given above (ideally always picking $\max_{k \in \mathcal{K}} \mathrm{R}(u, l, k)$). We assume that the ad-network already learned (through previous data purchase, or the market, or additional profiling) the exposure associated with keywords for all users. Then its capacity to pick $k$ judiciously depends on the fact that it can recognize for which user and for which location this ad will be shown in real time. This is precisely what our system permits. We denote the series of ad impressions by $(u_1, l_1), \dots, (u_M, l_M)$, and we say $i \in \mathcal{I}$ if the identities of $u_i$ and $l_i$ are disclosed to the ad-network and $i \notin \mathcal{I}$ otherwise (e.g. $u_i$ is anonymous at that location). Then its total revenue is:

$$\mathrm{R} = \sum_{i \in \mathcal{I}} \max_{k \in \mathcal{K}} \mathrm{R}(u_i, l_i, k) + \max_{k \in \mathcal{K}} \sum_{i \notin \mathcal{I}} \mathrm{R}(u_i, l_i, k)\,. \tag{5.2}$$

Note the interchange of the sum and max operator. As more information is purchased, $\mathcal{I}$ grows and terms from the right (where $k$ is chosen according to the "common denominator" of interests) are moved to the left (where $k$ is chosen to maximize each ads separately).

## Large-Scale Location Value Evaluation

After formulating this simple model, we use it to understand the "cost of privacy." Concretely, we want to understand what happens to revenue when companies are not able to use targeted advertisements at sensitive locations. To achieve this, we obtained large data sets of user location information, labeled each location with keywords, and estimated the monetary value of each keyword. We labeled the privacy sensitivity of each keyword according to sociological literature. For different categories of keywords, we viewed the effect on overall revenue.

**The overall revenue distribution** is calculated by applying Eq.(5.2) to our data set. Various values of $\delta_k$ were used without them affecting our results qualitatively; below we fix $\delta_k = 0.1$. We analyzed the distribution of revenue across users, locations, and keywords. Unsurprisingly, a low number of checkins resulted in low revenue. More interestingly, we found that locations and users with high numbers of checkins did not all have high revenue. This suggests some checkins can be kept private without greatly affecting overall revenue.

**The impact of privacy on revenue** can now be analyzed as follows. To first determine what keywords may be privacy sensitive, we used the work of J. Bing, who came up with a comprehensive classification of words and topics that can be taken to be privacy sensitive [7]. We found 33 blacklisted keywords in the Foursquare data set and 35 keywords in the CDR datasets[7].

We then studied the relationship between frequency, sensitivity, and value of keywords. Intuitively, a rare keyword (*i.e.*, one that is present in few locations or is visited only by a few users) may be more sensitive as it reveals more information. We wished to see if such keywords could be blacklisted without greatly impacting the advertising revenue.

As can be seen in Fig.5.3 (a), the CPC of a keyword does not seem to correlate in any way with its frequency represented on the graph in log-log plot. However, when we consider the total revenue from ads targeted to this keyword in Fig.5.3 (b), its frequency is quite positively correlated. Going a step further, when one considers the total revenue from all ads shown at a location associated with a specific keyword in Fig.5.3 (c), the correlation seems almost perfect. This result indicates that, fortunately, the keywords that are the most rare are also ones that generate little revenue and can be ignored, giving more evidence to a previous finding [106]. Note also that privacy-sensitive keywords appeared to be evenly distributed across CPC and revenue.

We next split the blacklist into several smaller blacklists based on categories. For each category, we calculated revenue per user without those blacklisted keywords. In other words, all locations that had blacklisted keywords associated with them were

[7] Example of blacklisted keywords: cannabis clinics, doctors, general litigation, gay bars, Buddhist temples. Examples of other keywords: home decor, painters, golf.

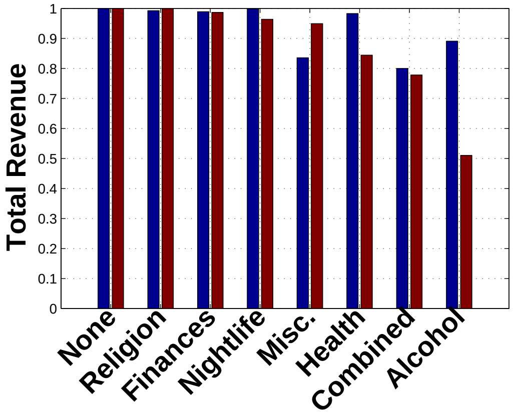


Figure 5.4: Effect of blacklisted keywords on revenue, per category. Foursquare (left bar), CDR (right bar)

*not* released when we calculated revenue. The drop in revenue when each specific category was blacklisted is shown in Fig. 5.4. We note that for locations related to religion and nightlife there was little to no drop in revenue for the Foursquare dataset (Fig. 5.4 (a)). This essentially means that these locations need *not* be released and yet there will be little perceptible drop in advertising revenue. For the CDR dataset (Fig. 5.4 (b)), we see religion and finance/legal to fall in this category. The 'combined' category refers to all categories except 'alcohol'. The alcohol category includes words like 'bar' and as such is highly conservative. We see that there is a perceptible drop in revenue when it comes to alcohol, but it is not more than 50% in the worst case. This points to a compromise position between privacy and advertising revenue.

## 5.5 Deployment and User Study

Having shown the system's robustness to several attacks and the high proportion of value retained when privacy sensitive locations are filtered out, we now present the specifics of the design. We also discuss an implementation and user study, conducted primarily to demonstrate the system's feasibility. The number of participants was too small for any broad conclusions, but we present results for completeness. IRB approval was granted for this study.

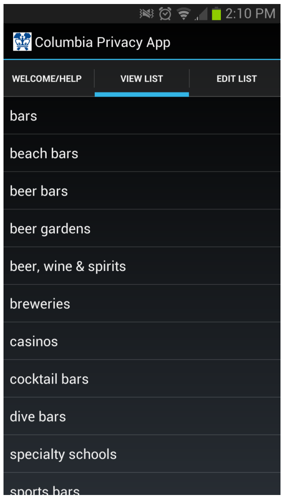
## Implementation

An implementation consists of several components: Software on the device, in order to hold a user's blacklist and report locations; a web server, to report keywords given a certain lat-long and store users' non-blacklisted locations; and a blocking module, to prevent information leakage. Our user study additionally included a web interface to publicly display all users' non-blacklisted locations.

We wrote our **location monitoring software** as an Android application due to Android's popularity. The app, available in Google Play[8], 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 these keywords into a 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 more likely to be considered sensitive (as defined in Sec. 5.4) 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.* The app passively recorded locations in the background every thirty minutes. Each half hour, the app would check the keywords in the current location. If any of these keywords were in the blacklist, no further action was taken. If none of the keywords were in the blacklist, the location and keywords would be uploaded to the server.

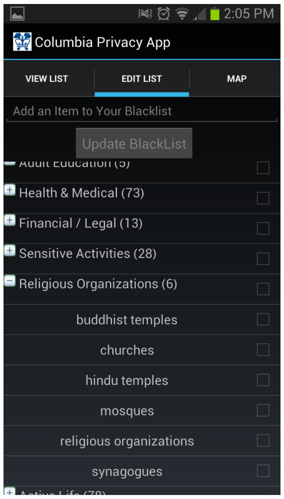
Our **webserver** had two main functions: reporting a location's keywords and storing a user's public location information. To map locations to keywords, we used

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



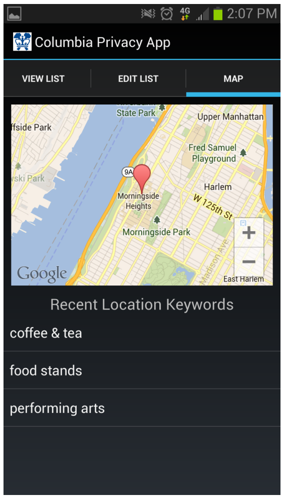
bars
beach bars
beer bars
beer gardens
beer, wine & spirits
breweries
casinos
cocktail bars
dive bars
specialty schools
sports bars

Figure 5.5: Blacklist



Adult Education (5) □
Health & Medical (73) □
Financial / Legal (13) □
Sensitive Activities (28) □
Religious Organizations (6) □
buddhist temples □
churches □
hindu temples □
mosques □
religious organizations □
synagogues □
Active Life (79)

Figure 5.6: Adding screen



Recent Location Keywords
coffee & tea
food stands
performing arts

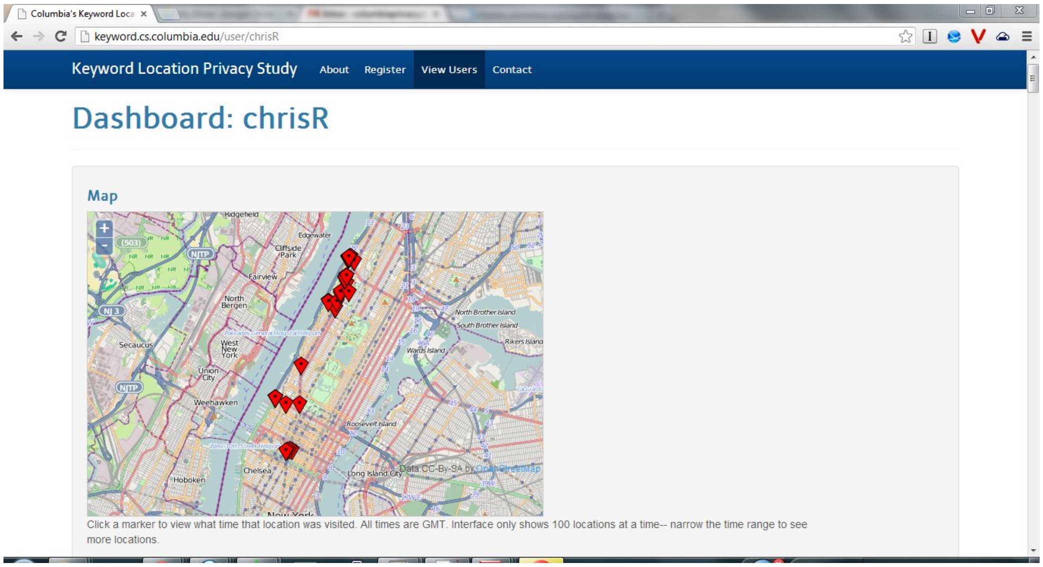
Figure 5.7: Map

the Yelp API. See Sec. 5.2 for more details on Yelp. Each time a user's device uploaded a lat-long to the server, we queried Yelp to find the categories of each business 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 on a user's privacy preferences. Yelp provides a hierarchical list of 885 categories and subcategories [9]. The server also stored all locations uploaded by the app in a database accessible only by the authors. In a full implementation, this server should additionally be able to communicate with ad exchanges.

For the purposes of our small scale user study, we did not think a **blocking module** was necessary. However, 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 in the middle and by 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[10]. In addition, as the system works with opt-in

[9] Yelp categories: `http://bit.ly/12TyTER`

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



Dashboard: chrisR

Click a marker to view what time that location was visited. All times are GMT. Interface only shows 100 locations at a time-- narrow the time range to see more locations.

Figure 5.8: Screenshot of our web interface (the data shown belong to an author, not a participant).

users, we can have the users 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.

The web interface, viewable at `keyword.cs.columbia.edu`, displayed all whitelisted locations, both on a map and listed with location keywords and times. In order to protect users' safety, we instructed users to contact us at any point if they were concerned about a misconfigured app resulting in unintentional location release. Additionally, any time a data point was recorded, we delayed making it public by 24 hours. Only users had the ability to see their data points in real-time via a password-secured link.

## Deployment

As this deployment was meant for exploratory purposes, we did not connect the system to any ad exchanges. We instead simulated the incentives and costs a user might experience while using our system. All participants received a small monetary sum for participating, and in addition 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. In order to simulate the costs of disclosing information, we publicly displayed a user's non-blacklisted locations through a web interface. 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 accurately 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.

We deployed our implementation with six users over roughly two weeks. The users were all living in America but were geographically diverse, including cities on both coasts and the Midwest. Study participants were recruited through advertising on social networks and were primarily adults in their mid-twenties.

### Observations

After completing the study, we asked users to complete a brief survey. Our study was too small to make any general conclusions, but we present results here to inform future work. Users expressed that they easily understood the keyword system and found the interface easy to use. The 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 claimed not to have used the blacklist at all. Others stated that they used the blacklist to hide sites they associated with social stigma or that they thought would send potentially negative signals to employers, insurers or the police.

## 5.6 Related Work

This chapter 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 [39, 42, 106, 115]. Privad [39] and Adnostic [115] 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 chapter is different – we are concerned with location information on mobile devices. Koi [42] 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. The work here is closely related to transaction privacy [106]. The difference is that we focus on location information for mobile devices and develop an economic model of location information to drive our market.

Bacelli et al [3] authors propose models to quantify the economic value of various locations, with the specific example of proximity advertising in mind. This is similar to our proposal, in that we too focus on proximity advertising and are interested in real-time location information. The main difference is that our approach is more empirically driven and much simpler with fewer assumptions. We rely on keywords associated with locations (derived from real data) and make no assumptions on how various businesses are distributed in a geographical region. We focus more on intent, captured by frequency of visits to a location, to approximate interest in a location and the propensity to conduct a commercial transaction at that location. Bacelli et al rely on a set of interests of a user that are known to the model.

## 5.7 Conclusion

Given the increasing ubiquity of mobile devices and the flourishing market of services for such devices, collection and monetization of location information has become a large concern. The main contribution of this chapter is the design of a solution that deals with location privacy using economics, and the analysis of this solution using both large data sets and a small deployment with real users.

Our solution is simple – opt-in users decide which locations they want to reveal and these revealed 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 location information that the user chooses not to release will not be leaked, and also minimizes the linkage of the user's identity with the released information.

We designed and analyzed the solution using Foursquare checkins, mobility patterns of millions of mobile subscribers from two large cities, business categories from an online review company, and CPC data from an ad-network. We find that potentially sensitive locations (as defined by sociological research) appear to be well distributed across locations sorted by popularity and profitability. Likewise, we find that the potential revenue of these sensitive locations is small compared to the total value generated from all locations. This suggests a sweet-spot between location privacy and monetizing location information. We construct and deploy a small scale version of the system with real users, showing that our solution is indeed feasible. We observe their behaviors and lay the groundwork for future study.

Chapter 6

# *Fair Location-Based Advertising*

In previous chapters, we've investigated anonymity, privacy, accountability, and transparency. We've designed tools to help individuals and shown how bias can exist at a group level. In this chapter, we set aside the problem of identifying a user and their private information and investigate the fairness of what happens with that location data. We examine the limits of fairness at both an individual and group level, presenting an empirical analysis of the impact of fairness on advertising revenue using a real world example: location based ad personalization for users of Instagram. We empirically analyze the potential for inadvertent discrimination among gender and race in location-based systems, additionally showing the impact of location representation on fairness. Furthermore, we apply fairness techniques to analyze how revenue is affected when both individual and group fairness guarantees must hold. Though the work in this chapter is a grounding for research into fairness in location-based ads, our methodology applies to more general advertising tasks.

## 6.1 Motivation and Summary of Results

We focus on informing what can *practically* be done to guarantee fairness when location data is used in targeted advertising. We choose this application for multiple reasons: It is increasingly common as location-based personalization reaches a large part of the population and it is hard to evade. As we empirically demonstrate, mobility data has great benefits but raises many concerns in the way it is currently used.

Perhaps more importantly, we show that many of the hardest challenges previously addressed in theoretical terms can be quantified in this scenario. For instance, this brings us to revisit questions like “What constitutes a practical definition of fairness?”, “What should we know or trust about those exploiting the data?”, “What is the gain we lose when some definition of fairness must be enforced?”

To start, we’ll describe a motivating example where disparate outcomes in targeted advertising is undesirable. For instance, consider a website advertising hiring opportunities to users; its goal is to optimize for relevance as long as disparate outcomes among genders and races are avoided. Why would such a system pose new challenges? First, previously proposed solutions focus on reconciling learning and fairness for *specific tasks for a single party* [134, 131, 19, 9]. For instance, how to increase loan repayment while satisfying equality of treatment or opportunity. In contrast, data providers interact with myriad third parties each leveraging data for different learning tasks. Second, as is commonly the case for online data providers, data about individuals are sparse and naturally represented in high dimensions. This contrasts with solutions designed to learn from a few structured features available for all users, such as exam scores. Additionally, leveraging data at large scale invariably means that computational complexity becomes a severe constraint, so each optimization to reconcile fairness with accuracy will rely on efficient approximation.

These challenges, however, do not imply that no solutions can be found to deploy fair targeting. The direction we examine here is to transform location data before they are used to train and target individuals. If the transformation and targeting satisfies some conditions (see background below), then fairness can be guaranteed for *any* task. As we demonstrate, much of the gains from targeting is preserved. For concreteness and simplicity, we focus in this short article on the simplest transform where details of mobile data are remove by grouping records into larger location cells.

## 6.2 Background

In our work, we use the definitions of “Fairness Through Awareness” [27], distinguishing between fairness at an individual level and at a group level, which we describe in detail below.

**Individual fairness.** The main principle is that similar people should see similar outcomes. More rigorously, we consider a classification setting where individuals (denoted by the set $V$) are mapped to probability distributions over outcomes $A$. For simplicity, throughout this work we will say each outcome is the decision of whether to show either a generic or targeted ad, and denote these outcomes as $A = \{0, 1\}$ with $A = 1$ corresponding to the decision to show a targeted ad and $A = 0$ a generic ad instead. The space of probability distributions defined on $A$ is $\Delta(A)$. From our point of view, a machine learning algorithm using data from the mobile ad-network defines a mapping $M : V \to \Delta(A)$. A difference score between individuals is denoted by $d : V \times V \to [0; 1]$ and a difference score between probability distributions is $D$. Throughout this chapter, without loss of generality, as a choice to measure the distance between probabilistic outcomes we will use $D_{TV}$, the distance of total variation (equivalent to one half the $\mathbb{L}_1$ norm) though others can be used. It is defined as: $D_{TV}(P, Q) = \frac{1}{2}\sum_{a \in A} |P(a) - Q(a)|$.

Given these definitions, an algorithm is *individually fair* if for all individuals $x$ and $y$, we have

$$D_{TV}(M(x), M(y)) \leq d(x, y) \tag{6.1}$$

Intuitively, this says that an advertising system must show similar sets of ads to similar users, and mathematically, this means that that an algorithm mapping users to distributions over outcomes must be Lipschitz continuous.

[27] shows that it is possible in polynomial time to find a mapping $M$ that is both individually fair and maximizes a linear objective function (such as expected revenue)

| Dataset | Number<br>Users | Number<br>Checkins | Labeled<br>Gender | Labeled<br>Race |
|---|---|---|---|---|
| New York | 22,300 | 707,265 | 10,388 | 902 |
| Los Angeles | 20,724 | 776,065 | 9,748 | 851 |

Table 6.1: Overview of dataset used in study.

using a linear program.

**Group fairness.** In contrast to individual fairness, [27] defines two groups of users $S$ and $T$ as having statistical parity up to bias $\varepsilon$ when:

$$D_{TV}\left(E_S[M], E_T[M]\right) \leq \varepsilon \tag{6.2}$$

where $E_S$ and $E_S$ denotes the expectation of ads seen by an individual chosen uniformly among $S$ and $T$. This definition implies that the difference in probability between two groups of seeing a particular ad will be bounded by $\varepsilon$. Note that individual fairness does not imply group fairness, and vice versa. A natural question is: "When can both individual fairness and statistical parity be achieved simultaneously"? To guide the design of a mobile platform one can use the following result introducing $d_{EM}(S,T)$, the Earth Mover's Distance [96] between $S$ and $T$.

## 6.3 Data Description

To understand the important trade offs facing advertising platforms, we collected a behavioral dataset linked to race and gender information, allowing us to study individual and group fairness and its impact on the predictive power of location data. We obtained publicly available data from Instagram, a popular image sharing social network and a valuable information source for several reasons. Instagram data includes behavioral data such as activity descriptions and locations – pieces of information linked to actions in the real world through the use of photos and smartphone GPS sensors. Additionally, photographs provide lots of user information, and the productization of computational vision techniques has made it possible to extract this information at scale.

## Methodology

We gathered 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. We then randomly sampled user profiles from those who had commented or liked his photos and gathered their metadata. We repeated this process, randomly sampling user IDs of those commenting or liking photos of any crawled profiles, obtaining the metadata of 115,796,284 for 260,389 different profiles. Systrom is a popular Instagram presence (7.9 million followers) and a wide variety of users comment on his photos, seemingly to communicate with the platform, making him a good starting point for a random crawl. No images were downloaded from Instagram.

**Location.** Of our 115 million photo information dataset, 16,537,404 were geotagged for 162,549 users. In order to study advertising that micro-targets small granularity locations, we narrowed our focus to two major United States cities, New York City and Los Angeles, a typical practice. Using only photos located in the bounding boxes of those two cities, we created two subsets: New York had 22,300 users with 707,265 photos and Los Angeles had 20,724 users with 776,065 photos.

**Tags.** Like other social networks, Instagram users label their content with "hashtags", which label topics for the photo, make photos more easily searchable, or let the user express him- or herself. As we discuss in a later section, we use these tags later as part of our location-based advertising model.

## Labeling

Next we discuss labeling our users with demographic information and evaluating that labeling.

**Labeling Gender.** To label our the gender of the users in our dataset, we applied the methodology of Mislove et al. [79]. We obtained the number of babies born by name, gender, and year of birth in the United States via Social Security data[1], assigning a gender to users with a first name for which there were both at least 50 births and 95% of recorded births were one gender. Out of our entire dataset of 260,389 users, this labeled 92,935 profiles (35%). In our New York City subset, 10,388 were labeled with gender, 5,471 female and 4,917 male. In Los Angeles, 9,748 users labeled with gender: 4,965 female and 4,783 male.

**Race labeling.** We labeled the race of profiles based on face recognition software, similar to prior work [90]. The Face++ API (`www.faceplusplus.com`) recognizes faces in images, additionally providing demographic information, labeling the race of users from one among Asian, Black, and Caucasian. Although we did not download any photos, our metadata included publicly accessible URLs of images, which we could pass to the Face++ API. We ran this software on the first 500 photographs of a subset of our New York and Los Angeles users, labeling a profile with the race that appeared most frequently in their photographs, using a binary labeling of Caucasian or minority. This labeled 902 users in our New York dataset; 746 labeled Caucasian and 156 from minorities, and 851 users in Los Angeles; 710 Caucasian and 141 minority.

**Evaluation with manual labeling**. To provide ground truth validation of our more scaled labeling techniques, two research assistants labeled a randomly selected subset of 200 profiles for gender and race. After filtering for private, deleted, or business profiles, 194 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.

Of our 194 human-labeled profiles, 86 users had first names recognized by our

[1] Available at `https://www.ssa.gov/oact/babynames/limits.html`

methodology. Of these, 84 out of 86 (97%) agreed, giving us high confidence in the precision of our gender labeling approach.

Comparing our race labeling methodology with the 194 human-labeled profiles and a binary Caucasian/Minority labeling, we found the image technique agreed with our human labelers 89.7% of the time, a lower level of accuracy than our gender labeling but still relatively high. This is in line with other works that report that Face++ has high levels of accuracy for race labeling.

## 6.4 Mobile Advertising Model

In order to analyze the trade off between fairness and revenue, we model a location-based advertising system using our dataset. We focus on this domain due to its importance (38% of all smartphone advertising used location targeting in 2016), and its potential for discrimination as location is highly sensitive and often correlates with sensitive traits such as race or income [104]. We simulate a system with the following problem: Given a user's locations from previous check-ins, predict what topics a user will be interested in. Such a prediction could allow a service to better target ads.

### User and Location Representation

We represent individuals in terms of their visits to different locations. We map locations to an index $j$. Each user is represented as an array, with index $j$ set to 1 if the user has checked in at location $j$ and a 0 otherwise. In our original dataset, locations for each photo are latitude-longitude pairs, and here we discretize these by truncating these coordinates to a certain level of prevision. In different analyses we vary this precision to study how fairness and revenue is impacted by granularity of location representation. Using fewer digits implies a lower granularity, which is better for privacy but less specific and hence likely less useful for advertisers. We vary the cell

| City | Race | Count | Fashion Interest | Health Interest | Travel Interest |
|---|---|---|---|---|---|
| NY | minority | 156 | 0.385 ± 0.039 | 0.141 ± 0.028 | 0.378 ± 0.039 |
| NY | white | 746 | 0.413 ± 0.018 | 0.155 ± 0.013 | 0.513 ± 0.018 |
| LA | minority | 141 | 0.298 ± 0.039 | 0.113 ± 0.027 | 0.355 ± 0.040 |
| LA | white | 710 | 0.332 ± 0.018 | 0.151 ± 0.013 | 0.420 ± 0.019 |

Table 6.2: Breakdown of tag interest by city and race. The percent of users who posted about the topic is shown with standard error of the mean.

sizes from 0 decimal places (*e.g.*, (-74., 40.) is a cell; cells have sides of length roughly 111km) to 4 places (e.g. (-73.9989, 40.7245) is a cell; cells have sides of roughly 10m). We additionally conducted our analysis representing users with a histogram of frequencies of visits to each location as opposed to binary representations, but the results were similar and we omit them due to space.

## Interest Prediction

After defining how users are represented, we use these feature to predict if a user is interested in several topics, utilizing Instagram's hashtags for ground truth. Hashtags, used on several platforms such as Instagram and Twitter, are ways for users to associate topics with their post. Examples include a user tagging a picture of food with "#food" or of himself with "#selfie". We use three different tags: #fashion, #travel, and #health.

We trained a model predicting a user's likelihood to post each of the three tags using a user's location visits as features and whether or not they had used a tag as labels. To avoid overfitting we regularized each model using ridge regression (i.e. $\mathbb{L}_2$ penalty) and conducting three way cross validation, picking the parameter that maximized peformance on the training set. All training was conducted using the scikit-learn python package.

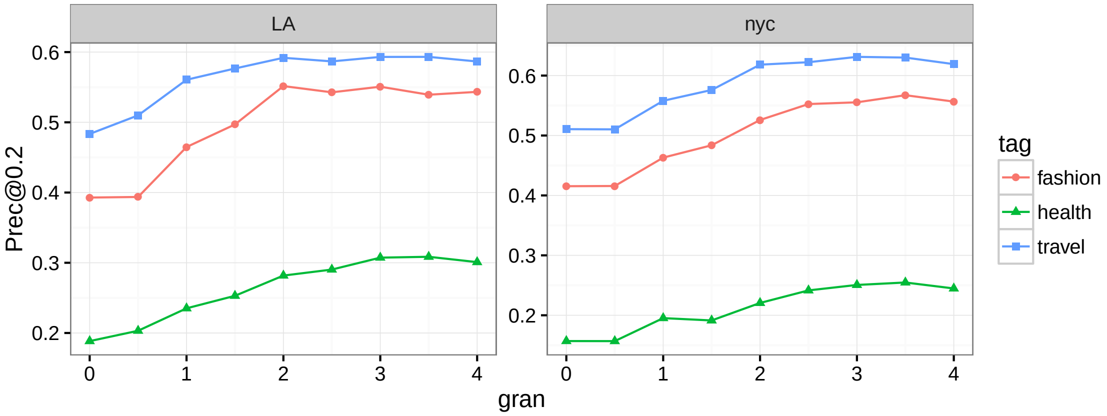


Figure 6.1: Granularity vs. Precision at 0.2 recall

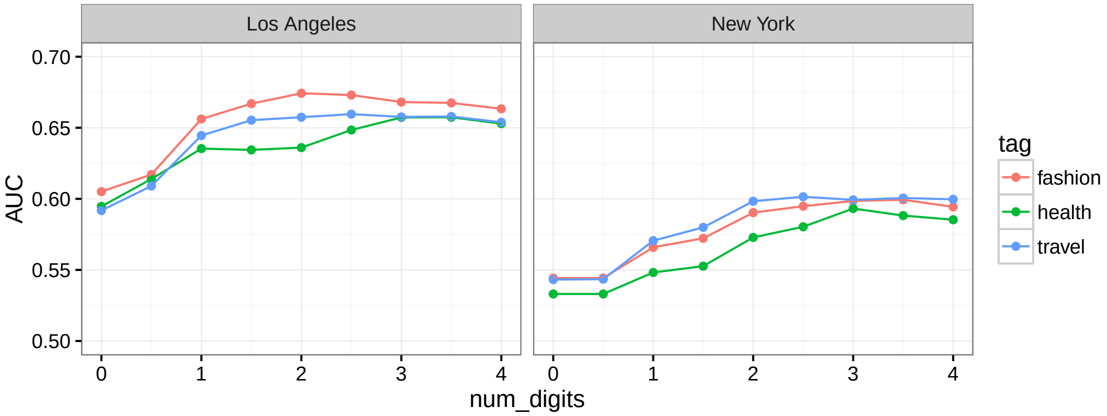


Figure 6.2: Granularity vs. AUC

## Performance and Revenue Estimation

We evaluate our models in two ways: in traditional machine learning terms and for their ability to improve revenue in an advertising simulation. We use AUC as a metric to understand our classifier performance due to its standard acceptance and our class distributions being highly skewed. For all three tags and both cities, AUC is 0.5 at the broadest granularity, meaning our model is no better than random guessing. However, as the number of digits increases, so does AUC. In NYC, our classifiers have AUCs of 0.82, 0.92, and 0.65 for fashion, health, and travel, respectively, and in LA, we report AUCS of 0.83, 0.92, and 0.68.

Moving beyond classifier performance, we estimated the impact of granularity

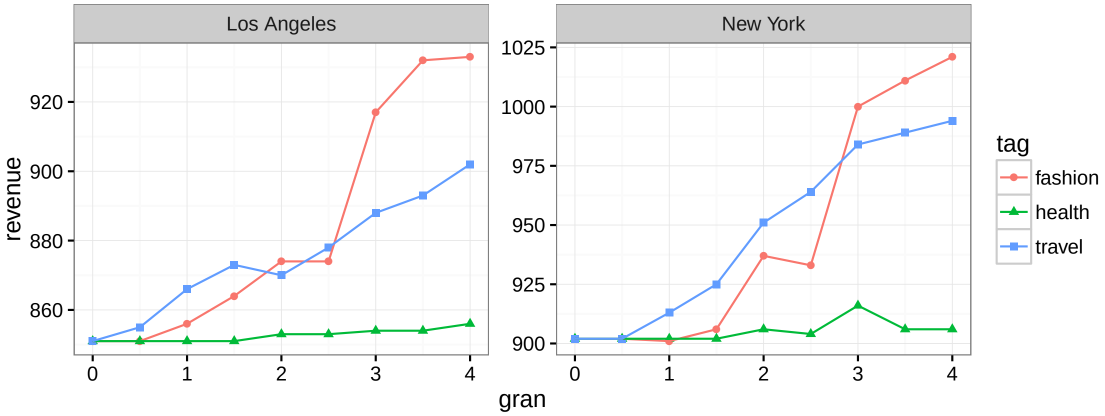


Figure 6.3: Revenue as a function of granularity, by city and tag.

on revenue. Earlier, we distinguished between generic and targeted advertisements. Based on estimates generated from the Facebook ad tool[2], we said that the cost per click (advertiser revenue) for a targeted ad was $2 and the revenue for a generic ad was $1. In our model, a generic ad always generates revenue, and a targeted ad only generates revenue if the user is indeed interested in a topic, and so the system will only show a targeted ad to a user if the expected revenue justifies the risk of receiving no revenue.

Figure 6.3 displays the impact of granularity upon revenue. The $x$ axis is latitude-longitude digits and the $y$ axis is revenue Point shape (and color) correspond to tags and each panel is a separate city. In New York, a predictor using the finest granularity of 4 digits generated $1021, $906, and $994 in revenue for fashion, health, and travel, respecitvely, over a baseline of displaying generic dislay $902. The optimal revenue if each interested user saw a targeted advertisement would be $1270, $1040, and $1344, meaning we achieved 71.0%, 86.7%, and 67.1%. The results were similar for LA with slight improvements on percent of optimal revenue.

[2] `https://www.facebook.com/business`

## 6.5 Evaluation

### Balancing Fairness and Revenue

We now consider revenue maximization under the constraint of individual fairness. In Section 6.2 we referenced how this could be achieved after the choice of a distance function between outcomes, a distance function between users, and a linear objective function. Our choice of $D$, the distance between distributions of ads, is $D_{TV}(P, Q) = \frac{1}{2}\sum_{a\in A}|P(a) - Q(a)|$. For our choice of $d$, the distance score between users, we again use the distance of total variation, this time upon the histogram of visits to locations between each pair of users using the representation of users defined in Section 6.4. Our objective function is to maximize expected revenue, as defined as $\sum_{x\in V} g\cdot\mu_x(0) + t\cdot\mu_x(1)$ with $g$, the revenue of a generic ad, set to 1 and $t$, the revenue of a targeted ad set, to 2. After these choices, the linear program chooses a probability of showing a targeted ad to a user to maximize revenue under the constraints of similar users seeing similar ads.

In order to make the trade-off between revenue and fairness more fluid, we differ from prior work and introduce a new parameter $k$ into Equation 6.1:

$$D_{TV}(M(x), M(y)) \leq k \cdot d(x, y) \tag{6.3}$$

A large $k$ means more flexibility in ad assignment but less individual fairness; $k = \infty$ means identical users can see completely different ads. In contrast, a low value of $k$ constrains the problem more, with $k = 0$ meaning all users must have the same ad distribution.

We run this linear program for both cities at all granularity levels and for multiple choices of $k$. We then compute a real revenue with the function $\sum_{x\in V} g\cdot\mu_x(0) + t\cdot\mu_x(1)\cdot\mathbf{1}_{x\in I}$ with the set $I$ denoting users who actually posted the target tag. Due to the number of constraints growing quadratically with the number of users, Here we are only able to present results for fairness by race and leave detailed analysis of gender for later work.

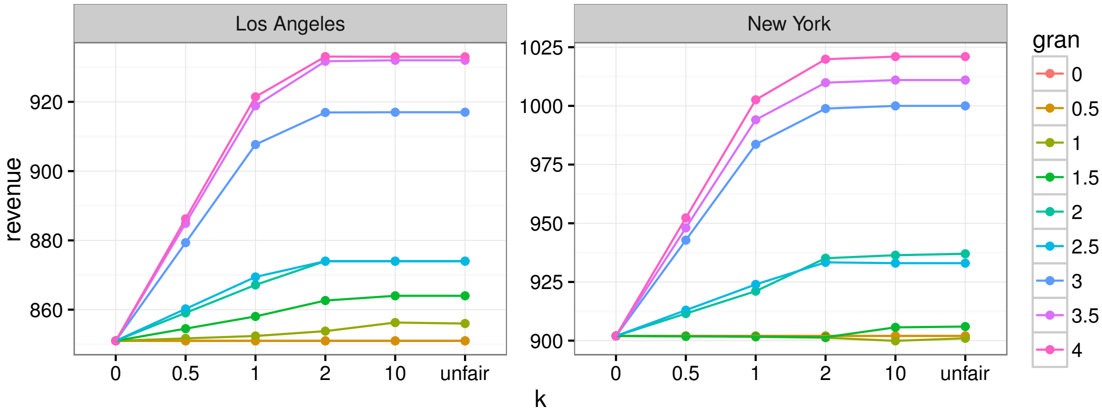


Figure 6.4: The impact of $k$ and granularity impact on revenue.

Figure 6.4 displays the impact of $k$ and granularity on revenue for both cities with the tag fashion. The $x$ axis corresponds to the choice of $k$ used in the linear program. The $y$ axis represents the actual revenue of the ad assignments output by the LP. Color denotes the granularity of location. The graph demonstrates again how finer granularity can increase revenue. In both NYC and LA, at nearly all values of $k$, a higher granularity corresponds to higher revenue. Another important takeaway is the shape of the lines. The revenue at $k = 2$ is nearly identical to the revenue at all higher amounts of $k$. The revenue declines rapidly at $k = 0$, where all individuals have the same distribution, and $k = 0.5$. The increase in revenue from $k = 1$ to higher values of $k$ is significant but not a large portion of the highest optimal revenue, suggesting a good potential value due to its balance and simplicity.

We next examine the impact of $k$ and granularity upon fairness. In Figure 6.4, the $x$ axis again corresponds to value of $k$. Color corresponds to race, with blue associated with caucasians and red associated with minorities. The $y$ axis now corresponds to the average probability that users of the class saw a targeted ad, with error bars corresponding to standard error of the mean. Each facet represents a different level of granularity.

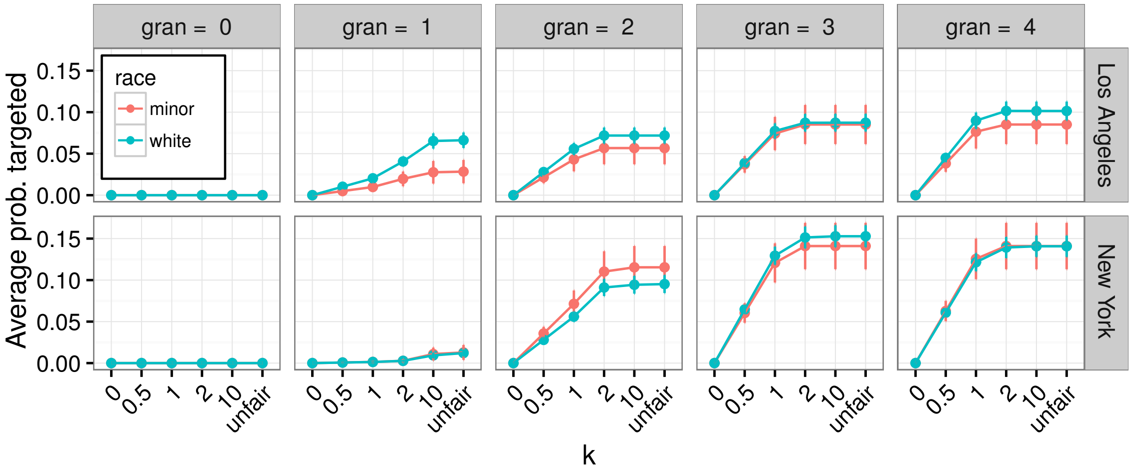


Figure 6.5: The impact of $k$ and granularity on fairness.

At lower levels of granularity, all users have similar low-resolution representations and thus it is difficult for our click predictor and then LP to risk displaying targeted ads, instead showing generic ads at all values of $k$. At medium level granularities, we see the algorithm begin to assign the ad to a small number of users and additionally the lines for each class to diverge, signally a rising level of group unfairness. Interestingly, in both graphs, the lines converge to be near identical at finer levels of granularity, at 4 digits for NYC and 3 and 3.5 digits for LA. This could be caused by mid-range granularities being associated more with neighborhoods, whereas very fine granularities will correspond to more exact venues, removing rougher associations of neighborhoods around areas with certain tags and narrowing them down to more specific places (e.g. 2 lat-long digits corresponds to roughly 1km, 4 to 10m).

## Bounding Fairness

For two demographic attributes, race and gender, we compute the Earth Mover's Distance, using the pyemd package [97, 96]. More precisely, for race we calculate the EMD between two probability distributions, one over Caucasian users and the other over Non-Caucasian users, with the "distance" between users defined as the distance of total variation of the histogram of their location visits. Similarly, for gender we calculate the EMD between the distribution of female and male users. As mentioned in Section 6.2 we represented locations as “cells”, assigning a photograph to a cell by truncating the latitude-longitude coordinates by a varying amount.

The large number of users labeled with gender presented a difficulty for our EMD calculation as Earth Mover’s Distance does not scale well. We use agglomerative clustering [123] to approximate EMD. We found this technique that groups individuals into “points” is well suited to our problem due to nonuniform cluster sizes.

We add a mechanism to cope with statistical parity, as it may create a spurious statistical bias between finite size groups, even when the expectations among those groups are equal. In addition to computing EMD between demographic groups, we also computed EMD between randomly created groups with the same size as our demographic groups.

In Fig. 6.6 we show the result of this process. The x-axis shows the granularity in terms of latitude longitude decimal places. The y axis shows the EMD. Lines are colored according to demographic, and a dashed line indicates random grouping of users as opposed to grouping by demographics. To put the EMD numbers into perspective, on the lower end, an EMD of 0.05 means one group may be seeing a targeted ad 5% more often. At the higher end of 0.8, users across the two groups are seeing quite different sets of ads.

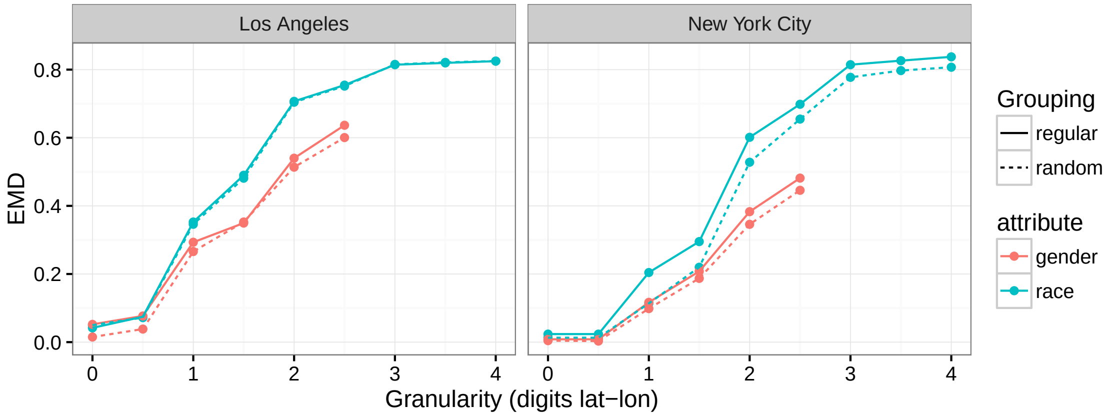


Figure 6.6: Risk vs. Granularity

In New York for race, the random line is clearly below the regular line, providing some evidence of real differences between the demographic groups as opposed to an artifact of sparsity. The line for gender is additionally more separate than it's counter-part in Los Angeles. This is possibly due to the much higher density in New York. As all users begin to have high difference scores from one another, caused by having no overlapping locations due to low density, all label assignments will be indistinguishable from each other. Gender overall seems to show a weaker separation between the real EMD and the random EMD.

The EMD increases as the data becomes more precise. One limitation of this study is that the distance $d$ we chose does not distinguish two users who have nearby but non-intersecting visits and users who are on the opposite side of the city. Different choices of $d$ with true geographical distance may refine those results.

## 6.6 Conclusion

In this chapter, we showed the impact of granularity on ad targeting, demonstrated the impact of fairness algorithms on a real world behavioral dataset, and explored a utility-fairness trade-off. There are many possible future directions. All results should be reproduced on larger datasets and different classes. One idea is to reformulate the problem in terms of *where* ads are shown or how users are reached, as opposed to focusing on the individuals. Building on our results, we also hope to create scalable algorithms for debiasing representations of users that work with sparse, large behavioral datasets.

# *Conclusion*

In this thesis, we have demonstrated both new problems and solutions in the field of location data. We have examined issues affecting both groups and individuals, and have considered solutions at both a local and system level. Throughout, we have have looked for flexible trade-offs as opposed to hard solutions which are likely to be opposed by data aggregators. Additionally, we have kept in mind the concepts of informed consent and user control as well as transparency.

In the first two chapters, we investigated attacks on user privacy. We showed that with just a few spatiotemporal points, it is possible to link anonymous users across anonymous datasets. In contrast to prior solutions which relied on heuristic techniques, our work was grounded in a theoretical model. We used novel datasets for evaluation that varied in behavior, location, and density. In the second chapter, we showed that knowledge of an individual's location data can be used to infer their race. We examined how mobility data can be used to create new metrics of segregation between a group and examined the trade off between granularity and classifier accuracy.

The later three chapters investigated solutions to the privacy-value trade off. Previous work on user privacy has often either completely blocked off user-level information from aggregators or has been of an abstract nature difficult for end-users to understand. In this portion of the thesis we focused on work that could empower users by informing them, would allow more choice in while balancing incentives of all system actors, and analyzed systems that prevented specific, understandable harms caused by data aggregation. In Chapter 4 we focused on transparency and accountability by building a website that helps users understand how their location data might be used to infer things about them. In Chapter 5 we proposed and analyzed a system whereby users broadly categorize the locations on which they can be appropriately targeted. This gave users an easy to understand method of data disclosure while putting locations in terms readily accessible to aggregators as well. The system was designed in a way such that users would be incentivized and indeed fairly compensated for their data. We showed that, assuming some reasonable user behavior, revenues would not be greatly negatively impacted by such a system. Finally, Chapter 6 combated algorithmic bias by empirically calculating the trade-offs between revenue, individual fairness, and group fairness. In contrast to previous studies which worked on specific problems or "dense" data, we explored a highly sparse and proposed a simple solution for de-biasing.

The work presented here answers some questions but creates many more. In this work, location is usually represented either as discretized geographic regions or as a collection of categories. Incorporating a better representation of location that takes into account both physical proximity as well as the semantics of the location could yield a greater benefit for location privacy or algorithmic bias research. Several of the works could benefit from questions of scaling: for example, in chapter 2, we could expand the work to more cities or countries, and in chapter 4 we could likewise scale our auditing website to more services.

One challenge that was heavily considered in this work and continues to face privacy research is this: do individuals care about privacy? People report "yes" when asked and the rise of encrypted chat applications and less broadcast-centric mediums like Snapchat seem to underline this. However, years of privacy research, data breaches, and privacy scandals have not impacted the fact that billions of people, including the author, trade their personal information for free services every day.

There are several possible explanations for such behavior. Maybe individuals don't actually care about privacy, but find it useful to report "yes" on surveys. Perhaps people do not understand the risks associated with the applications they use. Perhaps people understand the risks but are happy with the trade off, or perhaps the apps have become so indispensable as to allow little choice. In this thesis, we have endeavored to build systems and tools that are more easy for users to understand and utilize. Continued research, engineering, and communication with the public will be necessary.

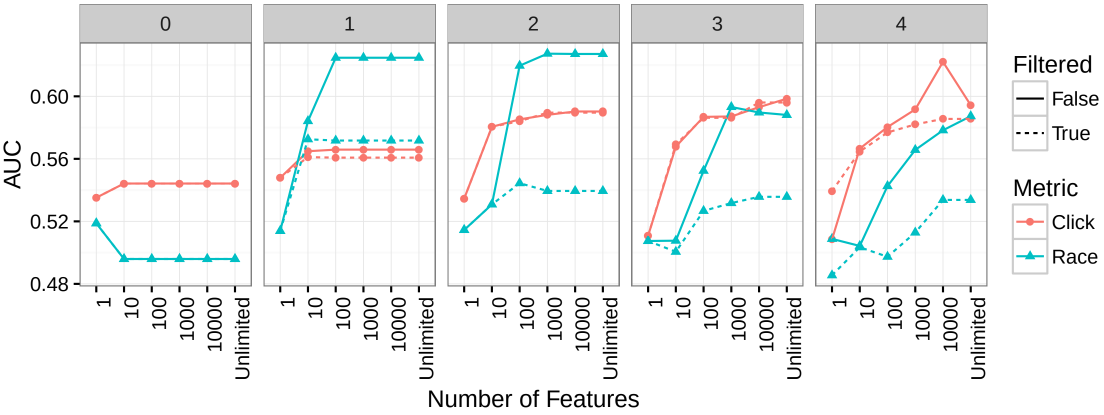


Figure 6.7: Removing features associated with race makes it difficult to predict race, but in this case does not greatly impact click prediction performance.

An interesting emerging area of study is finding "fair" representations of data [133]. What are scalable and interpretable ways to transform data, such that any subsequent machine learning or data-mining conducted on it will not show bias? As a follow up to the work in Chapter 6, I was curious about the performance of advertising classifier when possibly discriminatory features were removed. I used a simple approach: use a $\chi^2$ test for feature selection for a model that classifies race. I removed any features with a low $p$ value (under 0.1), and completed again the click prediction analysis of Chapter 6 and measure the performance. Additionally, I predicted race using the modified dataset. The results are displayed in Figure 6.7. The prediction results for clicks remain similar to as before, but now race is predicted with much lower performance. More investigation is required.

Data aggregation does not appear to be letting up, and is even growing into new domains of audio and video recording. Going forward, it will be very important to research and mitigate the negative impacts of large-scale data aggregation, in location data and in other forms. Doing so will necessarily be an interdisciplinary effort that includes many subjects beyond just privacy, such as security, economics, algorithmic bias, fairness, accountability, and transparency. This emerging field of study includes journalists who find real world harm, theoreticians defining the boundaries of what is possible, and dedicated engineers to bridge the gap between the two, bringing us the benefits of new, powerful insights into human behaviors without the concomitant concerns.

# *Bibliography*

[1] ACLU. *California Location Privacy Act of 2012*. 2012.

[2] Yaniv Altshuler et al. "Incremental learning with accuracy prediction of social and individual properties from mobile-phone data." In: *SocialCom/PASSAT*. IEEE. IEEE, 2012, pp. 969–974. ISBN: 978-1-4673-5638-1. URL: `http://dblp.uni-trier.de/db/conf/socialcom/socialcom2012.html{\#}AltshulerAFEP12`.

[3] Franois Baccelli and Jean Bolot. "Modeling the economic value of the location data of mobile users." In: *IEEE INFOCOM*. 2011, pp. 1467–1475.

[4] 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/. 2014.

[5] Mohsen Bayati et al. "Algorithms for Large, Sparse Network Alignment Problems." In: *Data Mining, 2009. ICDM '09. Ninth IEEE International Conference on* (2009), pp. 705–710.

[6] R.a. Richard Becker et al. "Human mobility characterization from cellular network data." In: *Communications of the ACM* 56.1 (2013), pp. 74–82. ISSN: 00010782. DOI: `10.1145/2398356.2398375`. URL: `http://dl.acm.org/citation.cfm?doid=2398356.2398375http://doi.acm.org/10.1145/2398356.2398375$\backslash$nhttp://neo-listas.udistrital.edu.co:2131/ft{\_}gateway.cfm?id=2398375{\&}type=pdf`.

[7] Joe Jon Bing. "Classification of Personal Information with respect to the sensitivity aspect." In: *Databanks and Society* (1972), pp. 98–150.

[8] Andrew J Blumberg and Peter Eckersley. "On locational privacy, and how to avoid losing it forever." In: (2009).

[9] Tolga Bolukbasi et al. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings." In: *Advances in Neural Information Processing Systems (NIPS)* (July 2016). arXiv: `1607.06520v1`.

[10] Laura Brandimarte et al. "Misplaced Confidences: Privacy and the Control Paradox." In: *WEIS* (2010).

[11] Jorge Brea et al. "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, 2014. ISBN: 9781450331920.

[12] J Candia, M González, and P Wang. "Uncovering individual and collective human dynamics from mobile phone records." In: *Journal of Physics A:* {...} (2008).

[13] A Cecaj, M Mamei, and N Bicocchi. "Re-identification of anonymized CDR datasets using social network data." In: *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on.* IEEE, 2014, pp. 237–242.

[14] Alket Cecaj, Marco Mamei, and Franco Zambonelli. "Re-identification and information fusion between anonymized CDR and social network data." In: *Journal of Ambient Intelligence and Humanized Computing* 7.1 (2015), pp. 1–14.

[15] Jonathan Chang et al. *ePluribus: Ethnicity on Social Networks.* 2010. URL: `https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1534http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewPDFInterstitial/1534/1828`.

[16] Ye Chen, Dmitry Pavlov, and John F Canny. "Behavioral Targeting." In: *ACM Transactions on Knowledge Discovery from Data* 4.4 (2010), pp. 1–31.

[17] Zhiyuan Cheng et al. "Exploring Millions of Footprints in Location Sharing Services." In: *Icwsm* 2010.Cholera (2011), pp. 81–88.

[18] Eunjoon Cho, Sa Seth A Myers, and Jure Leskovec. "Friendship and mobility: user movement in location-based social networks." In: *Proceedings of the 17th ACM SIGKDD ...* (2011), pp. 1082–1090. ISSN: 9781412946452. DOI: 10.1145/2020408.2020579. URL: `http://dl.acm.org/citation.cfm?id=2020579`.

[19] Alexandra Chouldechova. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments." In: *Proceedings of Workshop FATML* stat.AP (Oct. 2016).

[20] Peter Christen. *Data Matching, Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection.* Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[21] David J Crandall et al. “Inferring social ties from geographic coincidences.” In: *Proceedings of the National Academy of Sciences* 107.52 (2010), pp. 22436–22441. ISSN: 0027-8424. DOI: 10.1073/pnas.1006155107.

[22] Justin Cranshaw et al. “Bridging the gap between physical location and online social networks.” In: *Ubicomp ’10 Proceedings of the 12th ACM international conference on Ubiquitous computing* (2010), pp. 119–128. DOI: 10.1145/1864349.1864380. URL: http://portal.acm.org/citation.cfm?id=1864380.

[23] V. Dave et al. “Measuring and Fingerprinting Click-Spam in Ad Networks.” In: *ACM SIGCOMM.* 2012.

[24] Z. Deng and M. Ji. “Deriving Rules for Trip Purpose Identification from GPS Travel Survey Data and Land Use Data: A Machine Learning Approach.” In: *Traffic and Transportation Studies 2010.* 2010. Chap. 72, pp. 768–777. DOI: 10.1061/41123(383)73. eprint: http://ascelibrary.org/doi/pdf/10.1061/41123%28383%2973. URL: http://ascelibrary.org/doi/abs/10.1061/41123%28383%2973.

[25] M Duggan and J Brenner. “The demographics of social media users, 2012.” In: *Pew Research Center’s Internet {&} American Life Project* (2013).

[26] Cynthia Dwork, Aaron Roth, et al. “The algorithmic foundations of differential privacy.” In: *Foundations and Trends® in Theoretical Computer Science* 9.3–4 (2014), pp. 211–407.

[27] Cynthia Dwork et al. “Fairness through awareness.” In: *ITCS ’12 Proceedings of the 3rd conference on Innovations in Theoretical Computer Science.* 2012.

[28] William Enck and Others. “TaintDroid: an information-flow tracking system for realtime privacy monitoring on smartphones.” In: *USENIX OSDI.* 2010.

[29] C Kintana Et. al., Carmelo Kintana, and Others. “The Goals and Challenges of Click Fraud Penetration Testing Systems.” In: *ISSRE.* 2009.

[30] Thom File. *Computer and Internet Use in the United States.* http://www.census.gov/prod/2013pubs/p20-569.pdf. 2013.

[31] Al Franken, Sen. Al Franken, and Al Franken. *Location Privacy Protection Act.* 2011.

[32] O Goga et al. “Exploiting innocuous activity for correlating users across sites.” In: *WWW ’13: Proceedings of the 22nd international conference on World Wide Web*. 2013, pp. 447–458.

[33] Oana Goga et al. “On the Reliability of Profile Matching Across Large Online Social Networks.” In: *KDD ’15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ˜ACM Request Permissions, 2015, pp. 1799–1808.

[34] Lewis R Goldberg et al. “The international personality item pool and the future of public-domain personality measures.” In: *Journal of Research in personality* 40.1 (2006), pp. 84–96.

[35] Avi Goldfarb and Catherine E Tucker. “Online advertising, behavioral targeting, and privacy.” In: *Communications of the ACM* 54.5 (2011), p. 25.

[36] Marta C González et al. “Understanding individual human mobility patterns.” In: *Nature* 453.7196 (2008), pp. 779–82. ISSN: 1476-4687. DOI: 10 . 1038 / nature06958. arXiv: 0806 . 1256. URL: http://www.ncbi.nlm.nih.gov/pubmed/18528393.

[37] Google. “Google Location History.” In: *https://www.google.com/maps/timeline* (2015).

[38] Matthias Grossglauser and D. N C Tse. “Mobility increases the capacity of ad hoc wireless networks.” In: *Networking, IEEE/ACM Transactions on* 10.4 (2002), pp. 477–486. ISSN: 10636692. DOI: 10.1109/TNET.2002.801403.

[39] A Guha et al. “Verified Security for Browser Extensions.” In: *Security and Privacy (S{&}P), 2015 IEEE Symposium on* (2011), pp. 115–130.

[40] Saikat Guha, Bin Cheng, and Paul Francis. “Privad: practical privacy in online advertising.” In: *NSDI’11: Proceedings of the 8th USENIX conference on Networked systems design and implementation*. ˜USENIX Association, 2011.

[41] 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, 2012, p. 14. ISBN: 978-931971-92-8. URL: http://dl.acm.org/citation.cfm?id=2228298.2228317.

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

[43] H Haddadi. “Fighting Online Click-Fraud Using Bluff Ads.” In: *ACM CCR*. 2010, pp. 22–25.

[44] Aniko Hannak et al. “Measuring personalization of web search.” In: *WWW '13: Proceedings of the 22nd international conference on World Wide Web*. ˜International World Wide Web Conferences Steering Committee, 2013.

[45] Michaela Hardt and Suman Nath. “Privacy-aware personalization for mobile advertising.” In: *CCS '12: Proceedings of the 2012 ACM conference on Computer and communications security*. New York, New York, USA: ˜ACM Request Permissions, 2012, p. 662. ISBN: 9781450316514. DOI: 10.1145/2382196.2382266. URL: http://dl.acm.org/citation.cfm?doid=2382196.2382266.

[46] P Hornyack et al. “These aren't the droids you're looking for: retrofitting android to protect data from imperious applications.” In: *Proceedings of the 18th ACM conference on Computer and communications security* (2011), pp. 639–652.

[47] Yuheng Hu, Lydia Manikonda, and Subbarao Kambhampati. *What We Instagram: A First Analysis of Instagram Photo Content and User Types*. 2014. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8118.

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

[49] Sibren Isaacman et al. “A tale of two cities.” In: *HotMobile '10: Proceedings of the Eleventh Workshop on Mobile Computing Systems {&} Applications*. ˜ACM Request Permissions, 2010. ISBN: 9781450300056.

[50] Sibren Isaacman et al. “Identifying important places in people{\textquoteright}s lives from cellular network data.” In: *Pervasive Computing* (2011), pp. 133–151.

[51] Sibren Isaacman et al. “Ranges of human mobility in Los Angeles and New York.” In: *2011 IEEE International Conference on Pervasive Computing and Communications Workshops, PERCOM Workshops 2011* (2011), pp. 88–93. DOI: 10.1109/PERCOMW.2011.5766977.

[52] Shouling Ji et al. “Structure Based Data De-Anonymization of Social Networks and Mobility Traces.” In: *ISC Proceedings of the 17th International Information Security Conference*. Springer International Publishing, 2014, pp. 237–254.

[53] Ehsan Kazemi, Seyed Hamed Hassani, and Matthias Grossglauser. "Growing a Graph Matching from a Handful of Seeds." In: *Preprint* (2014), pp. 1–27.

[54] Patrick Gage Kelley et al. "When are users comfortable sharing locations with advertisers?" In: *CHI '11: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ~ACM Request Permissions, 2011, pp. 2449–2452. ISBN: 978-1-4503-0228-9. DOI: `10.1145/1978942.1979299`. URL: `http://portal.acm.org/citation.cfm?doid=1978942.1979299`.

[55] Kelton. *4th Annual SpringHill Suites Annual Travel Survey.* http://news.marriott.com/springhill-suites-annual-travel-survey.html. 2013.

[56] Nitish Korula and Silvio Lattanzi. "An efficient reconciliation algorithm for social networks." In: *Proceedings of VLDB* 7.5 (2014), pp. 377–388. arXiv: `1307.1690`. URL: `http://arxiv.org/abs/1307.1690`.

[57] D Koutra, Hanghang Tong, and D Lubensky. "BIG-ALIGN: Fast Bipartite Graph Alignment." In: *Data Mining (ICDM), 2013 IEEE 13th International Conference on.* 2013, pp. 389–398.

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

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

[60] John Krumm. "A survey of computational location privacy." In: *Personal and Ubiquitous Computing* 13.6 (2009), pp. 391–399.

[61] Mei-Po Kwan. "Gender, the Home-Work Link, and Space-Time Patterns of Nonemployment Activities." In: *Economic Geography* 75.4 (1999), pp–370.

[62] Neal Lathia, Daniele Quercia, and Jon Crowcroft. "The hidden image of the city: Sensing community well-being from urban mobility." In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 7319 LNCS (2012), pp. 91–98. ISSN: 03029743. DOI: `10.1007/978-3-642-31205-2_6`.

[63] Mathias Lécuyer et al. "XRay: Enhancing the Web{\textquoteright}s Transparency with Differential Correlation." In: *23rd USENIX Security Symposium (USENIX Security 14)* (2014).

[64] Ilias Leontiadis et al. "Don't kill my ads!: balancing privacy in an ad-supported mobile application market." In: *Proceedings of the Twelfth Work-*

*shop on Mobile Computing Systems {&} Applications* (2012), p. 2. DOI: 10.1145/2162081.2162084. URL: http://www.cl.cam.ac.uk/{~}il235/HotMobile12{_}Leontiadis.pdf.

[65] Kevin Lewis, Jason Kaufman, and Nicholas Christakis. "The Taste for Privacy: An Analysis of College Student Privacy Settings in an Online Social Network." In: *J. Computer-Mediated Communication* 14.1 (2008), pp. 79–100. URL: http://www.wjh.harvard.edu/{~}kmlewis/privacy.pdf.

[66] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity." In: *2007 IEEE 23rd International Conference on Data Engineering*. IEEE, 2007, pp. 106–115.

[67] L. Liao, D. Fox, and H. Kautz. "Extracting Places and Activities from GPS Traces Using Hierarchical Conditional Random Fields." In: *The International Journal of Robotics Research* 26.1 (2007), pp. 119–134.

[68] Kaisen Lin et al. "Energy-accuracy trade-off for continuous mobile device location." In: *ACM MobiSys*. 2010.

[69] Jack Lindamood et al. "Inferring Private Information Using Social Network Data." In: *Proceedings of the 18th International Conference on World Wide Web*. WWW '09. New York, NY, USA: ACM, 2009, pp. 1145–1146. ISBN: 978-1-60558-487-4. DOI: 10.1145/1526709.1526899. URL: http://doi.acm.org/10.1145/1526709.1526899.

[70] Bin Liu et al. "AdReveal: improving transparency into online targeted advertising." In: *HotNets-XII: Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks*. ˜ACM Request Permissions, 2013.

[71] Feng Liu et al. "Annotating Mobile Phone Location Data with Activity Purposes Using Machine Learning Algorithms." In: *Expert Syst. Appl.* 40.8 (2013), pp. 3299–3311. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2012.12.100. URL: http://linkinghub.elsevier.com/retrieve/pii/S0957417412013425http://dx.doi.org/10.1016/j.eswa.2012.12.100.

[72] M Madden. "Privacy management on social media sites." In: *Pew Research Center's Internet {&} American Life Project* (2012).

[73] Mary Madden et al. "Teens, Social Media, and Privacy." In: *Pew Research Center* (2013).

[74] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008. ISBN: 0521865719, 9780521865715.

[75] Douglas S Massey and Nancy A Denton. “The dimensions of residential segregation.” In: *Social Forces* 67.2 (1988), pp. 281–315.

[76] Dana Mattioli. “On Orbitz, Mac Users Steered to Pricier Hotels.” In: *online.wsj.com* (2012), pp. 1–6.

[77] 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: *The Melanin Millennium*. Ed. by Ronald E Hall. Dordrecht: Springer, 2013.

[78] Jakub Mikians et al. “Detecting price and search discrimination on the internet.” In: *HotNets-XI: Proceedings of the 11th ACM Workshop on Hot Topics in Networks*. July. New York, New York, USA: ˜ACM Request Permissions, 2012, pp. 79–84. ISBN: 9781450317764.

[79] Alan Mislove et al. “Understanding the Demographics of Twitter Users.” In: *ICWSM* (2011).

[80] Yves-Alexandre Y.-a. de Montjoye et al. “Unique in the shopping mall: on the reidentifiability of credit card metadata.” In: *Science* 347.6221 (2015), pp. 536–539. ISSN: 0036-8075. DOI: `10.1126/science.1256297`. URL: `http://www.sciencemag.org/content/347/6221/536.abstracthttp://www.sciencemag.org/content/347/6221/536`.

[81] Yves-Alexandre Y.-a. de Montjoye et al. “Unique in the shopping mall: on the reidentifiability of credit card metadata.” In: *Science* 347.6221 (2015), pp. 536–539. ISSN: 0036-8075. DOI: `10.1126/science.1256297`. URL: `http://www.sciencemag.org/content/347/6221/536.abstracthttp://www.sciencemag.org/content/347/6221/536`.

[82] Yves-Alexandre de Montjoye et al. “Predicting personality using novel mobile phone-based metrics.” In: *SBP’13 Proceedings of the 6th international conference on Social Computing, Behavioral-Cultural Modeling and Prediction* (2013), pp. 48–55.

[83] Yves-Alexandre de Montjoye et al. “Unique in the Crowd: The privacy bounds of human mobility.” In: *Scientific reports* 3 (2013), p. 1376. ISSN: 2045-2322. DOI: `10.1038/srep01376`. URL: `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3607247{\&}tool=pmcentrez{\&}rendertype=abstract`.

[84] Yves-Alexandre de Montjoye et al. “Unique in the Crowd: The privacy bounds of human mobility.” In: *Scientific reports* 3 (2013), p. 1376. ISSN: 2045-2322. DOI: `10.1038/srep01376`. URL: `http://www.pubmedcentral.`

nih.gov/articlerender.fcgi?artid=3607247{\&}tool=pmcentrez{\&}rendertype=abstract.

[85] Yves-Alexandre de Montjoye et al. “Unique in the Crowd: The privacy bounds of human mobility.” In: *Scientific reports* 3 (2013), p. 1376. ISSN: 2045-2322. DOI: 10.1038/srep01376. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3607247{\&}tool=pmcentrez{\&}rendertype=abstract.

[86] Move-O-Scope. “Move-O-Scope.” In: *https://move-o-scope.halftone.co/* (2015).

[87] Arvind Narayanan and Vitaly Shmatikov. “De-anonymizing Social Networks.” In: *2009 30th IEEE Symposium on Security and Privacy (SP)*. IEEE, 2009, pp. 173–187.

[88] Arvind Narayanan and Vitaly Shmatikov. “Robust De-anonymization of Large Sparse Datasets.” In: *Security and Privacy, 2008. SP 2008. IEEE Symposium on* (2008), pp. 111–125. ISSN: 1081-6011. DOI: 10.1109/SP.2008.33. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4531148.

[89] Arvind Narayanan et al. “Location Privacy via Private Proximity Testing.” In: *NDSS*. 2011.

[90] S Nilizadeh et al. “Twitter’s Glass Ceiling: The Effect of Perceived Gender on Online Visibility.” In: *Proceedings of the International Conference Weblogs and Social Media (ICWSM)* (2016).

[91] Anastasios Noulas et al. “An Empirical Study of Geographic User Activity Patterns in Foursquare.” In: *Fifth International AAAI Conference on Weblogs and Social Media* 11 (2010), pp. 570–573.

[92] Cathy O’Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2017.

[93] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. “Running experiments on amazon mechanical turk.” In: *Judgment and Decision Making* 5.5 (2010), pp. 411–419.

[94] Pedram Pedarsani and Matthias Grossglauser. “On the privacy of anonymized networks.” In: *KDD ’11: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ˜ACM Request Permissions, 2011, pp. 1235–1243.

[95] F Pedregosa and Others. "Scikit-learn: Machine Learning in {P}ython." In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[96] O Pele and M Werman. "Fast and robust earth mover's distances." In: *2009 IEEE 12th International Conference on Computer Vision* (2009).

[97] Ofir Pele and Michael Werman. "A linear time histogram metric for improved SIFT matching." In: *Proceeding ECCV '08 Proceedings of the 10th European Conference on Computer Vision.* Hebrew University of Jerusalem, Jerusalem, Israel. Dec. 2008, pp. 495–508.

[98] Marco Pennacchiotti and Ana-Maria Popescu. *A Machine Learning Approach to Twitter User Classification.* 2011. URL: `https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2886`.

[99] Delip Rao et al. "Classifying Latent User Attributes in Twitter." In: *Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents.* SMUC '10. New York, NY, USA: ACM, 2010, pp. 37–44. ISBN: 978-1-4503-0386-6. DOI: `10.1145/1871985.1871993`. URL: `http://doi.acm.org/10.1145/1871985.1871993http://portal.acm.org/citation.cfm?doid=1871985.1871993`.

[100] S F Reardon. "A Conceptual Framework for Measuring Segregation and its Association with Population Outcomes." In: *Methods in Social Epidemiology.* Ed. by J M Oakes and J S Kaufman. San Francisco, CA, USA: John Wiley Sons, 2006. Chap. 7, pp. 169–192.

[101] Chris Riederer and Augustin Chaintreau. "How You Know What They Know When They Know Where You ' ve Been." In: ().

[102] Christopher J. Riederer et al. "Challenges of keyword-based location disclosure." In: *WPES '13: Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society.* New York, New York, USA: ˜ACM Request Permissions, 2013, pp. 273–278. ISBN: 9781450324854. DOI: `10.1145/2517840.2517862`. URL: `http://dl.acm.org/citation.cfm?doid=2517840.2517862`.

[103] Christopher J Riederer et al. "I don't have a photograph, but you can have my footprints.: Revealing the Demographics of Location Data." In: *COSN '15: Proceedings of the third ACM conference on Online social networks.* ACM, 2015, pp. 185–195.

[104] Christopher J Riederer et al. "I don't have a photograph, but you can have my footprints.: Revealing the Demographics of Location Data." In: *Proceedings of the 2015 ACM on Conference on Online Social Networks.* ACM. 2015, pp. 185–195.

[105] Christopher Riederer et al. "Findyou: A personal location privacy auditing tool." In: *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee. 2016, pp. 243–246.

[106] Christopher Riederer et al. "For sale : your data: by : you." In: *HotNets-X: Proceedings of the 10th ACM Workshop on Hot Topics in Networks.* ˜ACM Request Permissions, 2011, pp. 1–6. ISBN: 9781450310598.

[107] 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." In: *Journal of the American Statistical Association* 66.336 (1971), pp. 755–759. URL: papers2://publication/uuid/5BA1A1D2-0AD0-4248-8B44-D82C82CD8AE3.

[108] Luca Rossi and Mirco Musolesi. "It{\textquoteright}s the Way you Check-in: Identifying Users in Location-Based Social Networks." In: *COSN '14: Proceedings of the 2nd ACM conference on Online social networks* (2014), pp. 215–226.

[109] 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*. 2014, pp. 836–843.

[110] Chaoming Song et al. "Limits of predictability in human mobility." In: *Science (New York, N.Y.)* 327.5968 (2010), pp. 1018–1021. ISSN: 1095-9203. DOI: 10.1126/science.1177170. URL: http://www.ncbi.nlm.nih.gov/pubmed/20167789.

[111] Mudhakar Srivatsa and Mike Hicks. "Deanonymizing Mobility Traces: Using Social Networks as a Side-Channel." In: *CCS '12: Proceedings of the 2012 ACM conference on Computer and communications security* (2012), pp. 628–637.

[112] Jacopo Staiano et al. "Money Walks: A Human-Centric Study on the Economics of Personal Mobile Data." In: *arXiv.org* (2014), p. 15. DOI: 10.1145/2632048.2632074. arXiv: 1407.0566. URL: http://arxiv.org/abs/1407.0566.

[113] 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.

[114] Latanya Sweeney. "k-anonymity: a model for protecting privacy." In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.5 (2002), pp. 557–570.

[115] V Toubiana, Arvind Narayanan, and D Boneh. "Adnostic: Privacy preserving targeted advertising." In: *Proc. NDSS* (2010).

[116] A Tuzhilin. *The Lane's Gifts v. Google Report.* 2006.

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

[118] "United States v. Jones." In: *S. Ct.* 132.10-1259 (2012), p. 945.

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

[120] Jayakrishnan Unnikrishnan. "Asymptotically Optimal Matching of Multiple Sequences to Source Distributions and Training Sequences." In: *Information Theory* 61.1 (2015), pp. 452–468.

[121] Jennifer Valentino-Devries, Jeremy Singer-Vine, and Ashkan Soltani. "Websites Vary Prices, Deals Based on Users' Information." In: *online.wsj.com* (2012), pp. 1–6.

[122] Wall Street Journal and Wall Street Journal. *Apple, Google collect User Data.* 2011.

[123] Joe H Ward. "Hierarchical Grouping to Optimize an Objective Function." In: *Journal of the American Statistical Association* 58.301 (Mar. 1963), pp. 236–244.

[124] Michael J White. "Segregation and diversity measures in population distribution." In: *Population Index* 52.2 (1986), pp. 198–221.

[125] C E Wills and C Tatar. "Understanding What They Do with What They Know." In: *WPES '12: Proceedings of the 12th annual ACM workshop on Privacy in the electronic society* (2012).

[126] Tim Wu. *The Attention Merchants: The Epic Scramble to Get Inside Our Heads.* Knopf, 2016.

[127] Xinyu Xing et al. “Exposing Inconsistent Web Search Results with Bobble.” In: *PAM ’14: Proceedings of the Passive and Active Measurements Conference.* 2014.

[128] Jun Yan et al. “How much can behavioral targeting help online advertising?” In: *WWW ’09: Proceedings of the 18th international conference on World wide web.* ˜ACM, 2009.

[129] Lyudmila Yartseva and Matthias Grossglauser. “On the performance of percolation graph matching.” In: *COSN ’15: Proceedings of the third ACM conference on Online social networks.* ˜ACM Request Permissions, 2013, pp. 119–130.

[130] Shuai Yuan et al. “Internet Advertising: An Interplay among Advertisers, Online Publishers, Ad Exchanges and Web Users.” In: *arXiv.org* (2012).

[131] Muhammad Bilal Zafar et al. “Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment.” In: *WWW ’17 Proceedings of the 26th International Conference on World Wide Web.* Oct. 2016.

[132] 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, 2011, pp. 145–156. ISBN: 9781450304924. DOI: 10.1145/2030613.2030630.

[133] Rich Zemel et al. “Learning fair representations.” In: *Proceedings of the 30th International Conference on Machine Learning (ICML-13).* 2013, pp. 325–333.

[134] Jiaming Zeng, Berk Ustun, and Cynthia Rudin. “Interpretable Classification Models for Recidivism Prediction.” In: *FATML* (Mar. 2015). arXiv: 1503.07810v6.

[135] 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.* ˜ACM Request Permissions, 2014, pp. 303–312.

[136] Yu Zheng. “Trajectory data mining: an overview.” In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 6.3 (2015), p. 29.

[137] Yuan Zhong et al. “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. New York, NY, USA: ACM, 2015, pp. 295–304.

[138] Yuan Zhong et al. "You Are Where You Go." In: *WSDM '15: Proceedings of the 8th ACM international conference on Web search and data mining*. ACM Press, 2015, pp. 295–304.

[139] statista.com. *Digital advertising spending worldwide from 2015 to 2020*. 2017. URL: `https://www.statista.com/statistics/237974/online-advertising-spending-worldwide/`.

*Appendices*

# A Proof of Theorem 1

We first show that each of the 2 factors in the denominator of $\phi(a_1, a_2)$ can be replaced by the corresponding truncated sum while affecting its value by at most $1 + 1/C^2$. Since the numerator is decreased by truncation, this establishes the upper bound on $\phi'(a_1, a_2)$. We then show that for the numerator of $\phi(a_1, a_2)$, the difference between the infinite sum and its truncated version is at most $1/C$ times the first term in this sum. Since the denominator is decreased by truncation, this establishes the lower bound on $\phi'$.

To obtain the upper bound, we first consider the factor $\sum_{k=a_1}^{\infty} \frac{\lambda^k}{k!}\binom{k}{a_1}(1-p_1)^{k-a_1}$ in the denominator. Expanding the binomial coefficient and pulling common terms outside the summation, this factor can be written as:

$$\frac{\lambda^{a_1}}{a_1!}\sum_{k \geq a_1} \frac{\lambda^{k-a_1}(1-p_1)^{k-a_1}}{(k-a_1)!} = \frac{\lambda^{a_1}}{a_1!}\sum_{k \geq 0} \frac{\lambda^{k}(1-p_1)^{k}}{k!}$$

Note that first term in this revised sum evaluates to 1, the term of index $\ln C$ evaluates to $\lambda^{\ln C}(1-p_1)^{\ln C}/(\ln C)! \ll \frac{1}{C^2}$, and the sum of all terms from $\ln C$ onward are at most $\frac{\lambda^{\ln C}(1-p_1)^{\ln C}/(\ln C)!}{(1-\lambda)}$ (upper bounding the infinite sum with a geometric series). Since $\lambda < 1/2$, we conclude that the sum of all terms from index $\ln C$ onward are less than $1/C^2$ times the first term.

The truncated sum for the second factor in the denominator can be bounded identically, giving us the desired upper bound on $\phi'(a_1, a_2)$.

It remains only to establish the lower bound by bounding the truncated numerator. We assume without loss of generality that $a_1 \geq a_2$. Expanding the binomial coefficients in the definition of the numerator of $\phi(a_1, a_2)$ and pulling common terms outside the summation, we can rewrite the numerator as:

$$\frac{\lambda_1^a(1-p_2)^{(a_1-a_2)}}{a_1!\ a_2!}\sum_{k \geq a_1} \frac{\lambda^{k-a_1}\left((1-p_1)(1-p_2)\right)^{k-a_1} \cdot k!}{(k-a_1)!(k-a_2)!}$$

The first term inside the revised sum is simply $a_1!/(a_1-a_2)! > 1$. Let $i$ denote the final index in the truncated sum, $a_1 + \max\{\ln C, 2a_1\}$. The $i$th term is upper bounded by $\lambda^{i-a_1} \cdot \frac{i!}{(i-a_1)!(i-a_2)!}$. If $a_1 \geq 4$, then since $i \geq 3a_1$, it is easy to see that $\frac{i!}{(i-a_1)!^2} < 1/2$. If $a_1 \leq 4$, then since $i - a_1 \geq \ln C \geq 7$ , we can note that $\frac{i!}{(i-a_1)!^2} < 1/2$. As $\lambda < 1/2$ and $i > a_1 + \ln C$, the $i$th term is less than $1/C \cdot 1/2$. Again upper bounding the infinite sum with a geometric series, the sum of all terms from index $i$ onward is less than the $i$th term divided by $(1-\lambda)$, and hence $< 1/C$. Therefore, the sum of all terms from the $i$th term onward is less than $1/C$ times the first term, completing the proof.

## Proof of Lemma 2

Recall that in Lemma 2, we proved that $E[\mathrm{Score}(u,v,\ell,t] \leq 0$ for any pair of users $u,v$ such that $v \neq \sigma_I(u)$. For $v = \sigma_I(u)$, we showed that the expected score is lower bounded by:

$$\begin{aligned}
&X(0,0)\ln\frac{X(0,0)}{Y(0,0)} + (1-X(0,0))\ln\frac{(1-X(0,0))}{(1-Y(0,0))}\\
&= X(0,0)\ln\frac{X(0,0)}{Y(0,0)} - (1-X(0,0))\ln\frac{(1-Y(0,0))}{(1-X(0,0))}\\
&\geq (1-\lambda(p_1+p_2-p_1p_2))\lambda p_1p_2 -\\
&\qquad \lambda(p_1+p_2-p_1p_2)\ln\frac{(1-e^{-\lambda(p1+p_2)})}{(1-e^{-\lambda(p_1+p_2-p_1p_2)})}
\end{aligned}$$

To prove that this expression is lower bounded by $(\lambda p_1p_2)^2K$, it suffices to prove that:

$$\begin{aligned}
&(1-\lambda(p_1+p_2-p_1p_2))\lambda p_1p_2 -\\
&\qquad \lambda(p_1+p_2-p_1p_2)\ln\frac{(1-e^{-\lambda(p1+p_2)})}{1-e^{-\lambda(p_1+p_2-p_1p_2)})}\\
&\geq (\lambda p_1p_2)^2K
\end{aligned}$$

or equivalently:

$$\begin{aligned}
&(1-\lambda(p_1+p_2-p_1p_2))p_1p_2 - \lambda(p_1p_2)^2K\\
&- \quad (p_1+p_2-p_1p_2)\ln\frac{(1-e^{-\lambda(p1+p_2)})}{(1-e^{-\lambda(p_1+p_2-p_1p_2)})} \geq 0
\end{aligned} \tag{4}$$

We can simplify the final factor in this inequality as follows:

$$\begin{aligned}
&\ln\frac{(1-e^{-\lambda(p1+p_2)})}{(1-e^{-\lambda(p_1+p_2-p_1p_2)})} = \ln e^{-\lambda(p_1p_2)}\frac{(e^{\lambda(p_1+p_2)}-1)}{(e^{\lambda(p_1+p_2-p_1p_2)}-1)}\\
&= \left(\ln\frac{(e^{\lambda(p_1+p_2)}-1)}{(e^{\lambda(p_1+p_2-p_1p_2)}-1)}\right) - \lambda p_1p_2
\end{aligned}$$

where the first equality came from multiplying the numerator and denominator by $e^{\lambda(p_1+p_2-p_1p_2)}$.

Substituting into Inequality (4), our lemma reduces to:

$$\begin{aligned}
&(1-\lambda(p_1+p_2-p_1p_2))p_1p_2 - \lambda(p_1p_2)^2K\\
&\quad (p_1+p_2-p_1p_2)\left(\ln\frac{(e^{\lambda(p_1+p_2)}-1)}{(e^{\lambda(p_1+p_2-p_1p_2)}-1}) - \lambda p_1p_2\right) \geq 0
\end{aligned}$$

or, equivalently:

$$\begin{aligned} & p_1p_2(1-\lambda(p_1p_2)K) - \\ & \qquad (p_1+p_2-p_1p_2)\ln\frac{(e^{\lambda(p_1+p_2)}-1)}{(e^{\lambda(p_1+p_2-p_1p_2)}-1)} \geq 0 \end{aligned} \tag{5}$$

This is hard to simplify directly, so we introduce the following upper bound:

$$\lambda p_1p_2 = \ln\frac{1}{e^{-\lambda p_1p_2}} = \ln\frac{e^{\lambda(p_1+p_2)}}{e^{\lambda(p_1+p_2-p_1p_2)}} \leq \ln\frac{e^{\lambda(p_1+p_2)}-1}{e^{\lambda(p_1+p_2-p_1p_2)}-1}$$

Using $Z$ to represent the quantity $\ln\frac{e^{\lambda(p_1+p_2)}-1}{e^{\lambda(p_1+p_2-p_1p_2)}-1}$ and substituting the new inequality in Inequality (5), we are trying to prove:

$$\begin{aligned} & p_1p_2(1-ZK) - (p_1+p_2-p_1p_2)Z \geq 0 \\ & \Leftrightarrow p_1p_2 \geq (p_1+p_2-p_1p_2(1-K))Z \\ & \Leftrightarrow \frac{p_1p_2}{p_1+p_2-p_1p_2(1-K)} \geq Z \\ & \Leftrightarrow e^{\frac{p_1p_2}{p_1+p_2-p_1p_2(1-K)}} \geq \frac{e^{\lambda(p_1+p_2)}-1}{e^{\lambda(p_1+p_2-p_1p_2)}-1} \end{aligned}$$

Now to conclude the proof we use two inequalities that follows from the Taylor expansions. In particular we have:

$$e^x \geq 1 + x + \frac{1}{2}x^2$$

and for $x \in o(1)$:

$$e^x \leq 1 + x + x^2$$

Now by assuming that $\lambda \in o(1)$ and by fixing $K = \frac{1}{2}\lambda(p_1 + p_2 - p_1p_2)^2$ we get:

$$\begin{aligned}
& e^{\frac{p_1p_2}{p_1+p_2-p_1p_2(1-K)}} \geq \frac{e^{\lambda(p_1+p_2)} - 1}{e^{\lambda(p_1+p_2-p_1p_2)} - 1} \\
\Leftrightarrow \quad & 1 + \frac{p_1p_2}{p_1 + p_2 - p_1p_2 + \frac{1}{2}\lambda(p_1 + p_2 - p_1p_2)^2} + \\
& \frac{p_1^2p_2^2}{2(p_1 + p_2 - p_1p_2 + \frac{1}{2}\lambda(p_1 + p_2 - p_1p_2)^2)^2} \geq \\
& \frac{\lambda(p_1 + p_2) + \lambda^2(p_1 + p_2)^2}{\lambda(p_1 + p_2 - p_1p_2 + \frac{1}{2}\lambda(p_1 + p_2 - p_1p_2)^2)} \\
\Leftrightarrow \quad & 1 + \frac{p_1p_2}{p_1 + p_2 - p_1p_2 + \frac{1}{2}\lambda(p_1 + p_2 - p_1p_2)^2} + \\
& \frac{p_1^2p_2^2}{2(p_1 + p_2 - p_1p_2 + \frac{1}{2}\lambda(p_1 + p_2 - p_1p_2)^2)^2} \geq \\
& 1 + \frac{p_1p_2 + \lambda(p_1 + p_2)^2}{p_1 + p_2 - p_1p_2 + \frac{1}{2}\lambda(p_1 + p_2 - p_1p_2)^2} \\
\Leftrightarrow \quad & \frac{\frac{1}{2}p_1^2p_2^2}{p_1 + p_2 - p_1p_2 + \frac{1}{2}\lambda(p_1 + p_2 - p_1p_2)^2} \geq \lambda(p_1 + p_2)^2
\end{aligned}$$

Now by fixing $\lambda < \frac{1}{8}\frac{p_1^2p_2^2}{(p_1+p_2)^2}$ we get:

$$\begin{aligned}
& \frac{\frac{1}{2}p_1^2p_2^2}{p_1 + p_2 - p_1p_2 + \frac{1}{2}\lambda(p_1 + p_2 - p_1p_2)^2} \geq \lambda(p_1 + p_2)^2 \\
\Leftrightarrow \quad & \frac{\frac{1}{2}p_1^2p_2^2}{p_1 + p_2 - p_1p_2 + \frac{1}{16}p_1^2p_2^2} \geq \frac{1}{8}p_1^2p_2^2 \\
\Leftrightarrow \quad & \frac{1}{4}p_1^2p_2^2 \geq \frac{1}{8}p_1^2p_2^2
\end{aligned}$$

So the claim follows.

# B Model of Location Value

Our economic model for location information leverages keywords that can be related to physical locations. At the same time, much like keywords used by ad-networks, they characterize the typical demands for impressions on this given topic. The premise is that a location has high economic value if keywords associated with the location have corresponding high value, as given by ad-networks and aggregators.

## Brief overview of ad networks

Online advertising is the *raison d'être* of collection of personal information about mobile users. This complex ecosystem involves three kinds of actors: the advertisers who wish to promote their products, the publishers who are in the context of this paper the app developers creating ad impressions that can be monetized, and the users who access the apps and services that publishers create. The ad-network is the entity responsible to orchestrate the interactions among these players, maximizing the revenue through better matching between products, ad-context, and users.

**Keywords** are pervasive in online advertising. They are used in sponsored search, as well as display ads, to interpret the context under which an ad is placed. They are also used to characterize the behavior of a user (previous search queries [128], or terms found in browsing history [16]). Advertisers use keywords to decide where to show their products.

Multiple revenue models exist that share differently the uncertainty associated with an advertisment. In a cost-per-mille (CPM) model, the publisher and ad-network receive a constant price for all the ads they show, hence the risk is entirely taken by the advertisers. Although we do not model this case here, our solution can be applied to it. In fact information about users is already sold through market of third party cookies like `bluekai.com`.

**The cost-per-click (CPC) model**, which is the one we analyze here, implies that the ad-network takes most of the risk as it will be paid only when users react to the impression by clicking on the ad. Advertisers hence continue to bid for clicks associated with certain keywords, as they do today. We also assume that this keyword automatically define a set of places that are relevant for this advertisement to be effective. The advertiser could also specify manually a *target* set of locations, a feature we do not model here but that would immediately fit in our solution. An ad network receives all keyword bids. Its ability to match product and users is the primary reason why additional information about the user is critical. Contextual advertising, and Behavioral targeting are two common techniques used to extract additional revenue when such information is available. The ad-network in this case needs not only to decide which ad to placed based on which advertisers bids the highest amount per click, but it also need to estimate the chance that a click occurs when a specific ad is played. We adopt a simple linear model based on a user exposure to keywords to model this choice.

More details on online advertising can be found in [130].

## Places, Keywords, Mobility

We model the mobility of users using a simple discrete model of visits to a set of pre-defined *locations* $\mathcal{L}$ that we index using $l$. A location may denote a point of interest where users check in, as in location based services like Foursquare. A location may alternatively represent a certain geographical area defined either using longitude/latitude coordinates, or the coverage of a given cell tower. Our model will be evaluated for both cases later.

A location $l$ implicitly provides information about the users who visit it (*i.e.*, people visiting a pet-shop are implicitly expressing an interest in pets). Our model represents this as follows: we associate a set of *keywords* to each location, that correspond to categories or topics of interest that are relevant to this particular location. Note that a single location can have multiple keywords and, similarly, that a keyword may be present in multiple locations, especially a popular one such as “coffee” or “music”. Keywords in our model are indexed by $k \in \mathcal{K}$. The relation between locations and keywords may be thought of a bipartite graph and we define the association matrix $\mathbb{A}$ as:

$$A_{l,k} = 1 \text{ if } k \text{ is a keyword for } l, 0 \text{ otherwise.}$$

The mobility of each user $u \in \mathcal{U}$ can be modeled as a random jump process $Y_u(t)$ of locations taking value in $\mathcal{L}$. Given that human mobility is periodical, it makes sense to consider the intensity of visit of a user at a location, denoted by $\mu_u(l)$. This variable captures the rate per unit of time that this location is visited, or equivalently the fraction of time spent at it.

### Exposure

A user throughout the day visits several locations. Some may visit more frequently locations in which particular keywords are present. The following variable called *exposure*, measures the fraction of time a user spends in a location relevant to this topic:

$$X_u(k) = \frac{\sum_{l \in \mathcal{L}} A_{l,k} \mu_u(l)}{\sum_{l \in \mathcal{L}} \mu_u(l)}\,.$$

A location containing fewer keywords implicitly provides more specific information about the users who visit it. It is hence important to consider the *normalized exposure*:

$$\tilde{X}_u(k) = \frac{\sum_{l \in \mathcal{L}} \frac{1}{\text{supp}(l)} A_{l,k} \mu_u(l)}{\sum_{l \in \mathcal{L}} \mu_u(l)}\,, \text{ with } \text{supp}(l) = \sum_{k \in \mathcal{K}} A_{l,k}$$

Normalized exposure is likely to be correlated with the intrinsic interest of a user in this particular topic. We should hence expect users to react more positively to ad on topics they have been more exposed to.

# C Appendix: Blacklists

### CDR Blacklist

- Acupuncture
- Adult
- Adult Entertainment
- Buddhist Temples
- Cannabis Clinics
- Casinos
- Chiropractors
- Cosmetic Surgeons
- Counseling & Mental Health

- Dentists
- Dermatologists
- Doctors
- Endodontists
- Family Practice
- Financial Advising
- Gay Bars
- General Litigation
- Health and Medical
- Home Health Care
- Hookah Bars
- Laser Eye Surgery/Lasik
- Lawyers
- Lingerie
- Maternity Wear
- Medical Spas
- Naturopathic/Holistic
- Obstetricians and Gynecologists
- Ophthalmologists
- Optometrists
- Oral surgeons
- Orthodontists
- Osteopathic physicians
- Pawn shops
- Pediatric dentists
- Police departments
- Religious organizations
- Traditional chinese medicine
- Urgent care
- Weight loss centers

## Foursquare Blacklist

- Assisted Living
- Bank
- Campaign Office
- Capitol Building
- Casino
- Cemetery
- Church
- City Hall
- Cosmetics Shop
- Courthouse
- Credit Union
- Dentist’s Office

- Doctor's Office
- Drugstore /Pharmacy
- Emergency Room
- Financial or Legal Service
- Fire Station
- Funeral Home
- Gay Bar
- Government Building
- Home (private)
- Hospital
- Lingerie Store
- Medical Center
- Middle School
- Military Base
- Mosque
- Playground
- Police Station
- Racetrack
- Strip Club
- Synagogue
- Tattoo Parlor