Big Location Data:
Balancing Profits, Promise, and Perils

Thesis Proposal

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

February 15, 2017

Abstract

Ubiquitous, mobile computing in the form of smartphones has created data that lets us study human behavior like never before. In particular, data about human mobility has allowed us to understand the hows and whys of human movement. At the same time, these new collections of data can present societal risks, as we've now enabled mass surveillance, a loss of privacy, and algorithmic bias.

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

Contents

1 Introduction

TODO

1.1 Outline

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

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

Chapter 4 examines the possibility of anonymizing location data. Prior work has shown that users are highly unique in their location patterns, leaving them vulnerable to deanonymization (see 2.2). Here we take this a step further, showing not only that this vulnerability exists, but that users indeed can be linked to other datasets. Additionally, we provide a tool to users that aggregates and displays their location data along with the potential inferences made from it.

Chapter 5 shows the potential for location data to be part of systems that We gather a dataset of locations attached to demographic information from a popular image-sharing mobile application. This data allows us to study the differences in human mobility across different groups, and moreover, to show that demographics can be inferred using only location data. This raises questions about the sensitivity of location data, and about the potential for bias in systems that make decisions based on location data. We examine other methodologies for inferring demographics from social network data and discuss debiasing of algorithms.

Chapter 6

Chapter 7

I conclude with a plan for completing this work in Chapter 8.

2 Background

2.1 Location Data

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

Location data can be described in two main ways: **geographically** or **semantically**. *Geographic* data can be described by a latitude-longitude data on the globe. *Semantic* location data refers to an identifier used within that dataset. This could have some information available to a

common user, e.g. "New York City", or it could simply be an identifier, e.g. 7. Note that often these two may be combined or used together. A location such as "CEPSR Office 618, Columbia University" (the author's office) indicates a very small, non-ambiguous location that can easily be mapped to geographic coordinates. Semantic location data can sometimes present a privacy problem, as an association with a place could indicate sensitive attributes, such as someone's religion, political affiliation, health, or sexuality. In this work, I will typically assume location data is also tagged with temporal data, and I will use the terms location data and spatiotemporal data interchangeably.

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

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

or, including time:

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

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

How is location data collected? Location data can be captured passively or actively. **Actively captured** location data is only recorded when the user takes some action. Note that this action does not need to inherently be "about" location data, for example, a user making a call from a cell phone or swiping a credit card is typically not consciously thinking about their location data. A record of their location is created as a by-product of their use of that technology. **Passively captured** is meant in a stronger way– the user's location is captured without the user making any kind of action. This can occur through tracking apps. An example is MapMyRun[1], an app where users record their routes while running, in order to track distance and progress in meeting exercise goals. Although the user took an action to start recording their location, the location is recorded in the background with no user action from then on, and hence we call it "passive". Another example is Google's location history. Google records location data in the background of a users Android phone every few minutes. A map of everywhere a user (with an Android phone with location history turned on) is available at [2].

What is location used for? TODO

2.2 Privacy

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

ADD MORE STUFF

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

k-anonymity

Differential privacy

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

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

2.3 Bias

2.4 Online Advertising

3 Location Data, Privacy, and Economics

The online economy is based primarily on advertising. The income of a firm roughly translates to (number of impressions) x (dollars per impression). I am trying to keep this abstract and not saying that firms are always getting paid for impressions, as other models like paying per click or per sale or other action are quite common. Really the argument here is that firms make money based on how many people come to their site and how well they can target advertisements to those individuals. This gives firms an incentive to gather as much information about their users as possible so that they can better target ads to them.

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

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

Therefore it is important to gain an understanding of how users value their information, what they believe firms are doing with their data, and what users are comfortable with in terms of data use.

3.1 Related Work

3.2 Completed Work

3.2.1 Your Browsing Behavior for a Big Mac

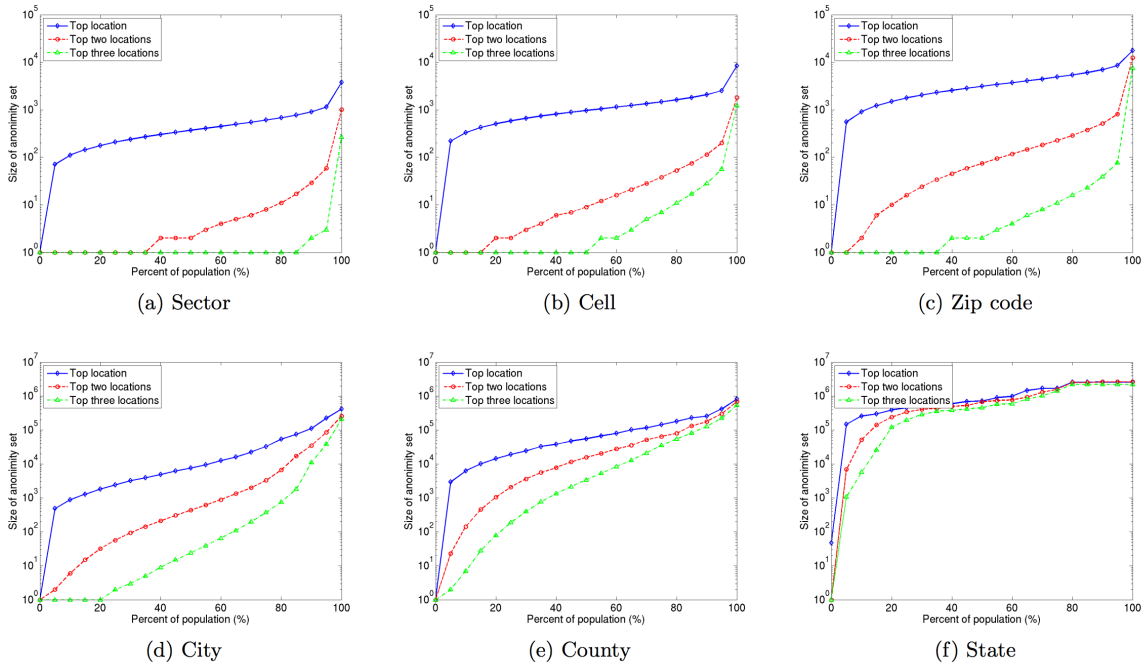
How does one determine how a study participant values something as abstract as User privacy is extremely important. However, there does not exist a strong understanding of how users value their privacy.

3.2.2 For Sale: Your Data. By: You

[1]

3.2.3 Challenges of Keyword-Based Location Disclosure

[3]



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

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

Figure 1: Figure from [4] 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.

4 Location Data and Anonymity

4.1 Related Work

Location data for individuals is highly unique and thus difficult to anonymize. The first large-scale study of the k-anonymity of location data was appropriately titled "Anonymization of Location Data Does Not Work" [4]. The paper used data from cell phone call detail records (or CDR, see Chapter 2) for 25 million United States users over a 3 month period. The authors represents each user as simply their top n most visited locations, varying n from 1 to 3. Additionally, the authors varied the granularity of the locations, with the smallest as cell sector and the largest as state. Remarkably, using 3 locations at a cell level made half of all users completely unique, and 3 locations a sector level made 85% of all users unique. A figure detailing this result and results for other granularities and values of n is depicted in Figure 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.

The Montjoye nature report

4.2 Completed Work

I have investigated the anonymity of location data for users

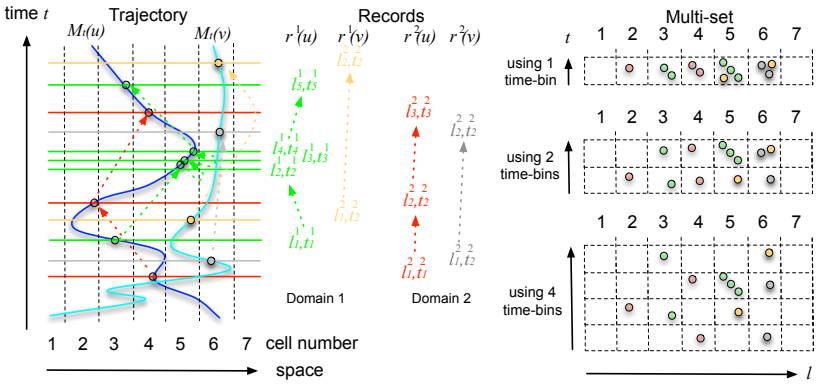


Figure 2: Two space-time trajectories with associated footprints in two domains.

4.2.1 Linking Users Across Domains with Location Data

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

TODO: put some math here?

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

TODO: put some math here?

In "Linking Users Across Domains with Location Data", published at WWW in 2016 [2], we tackled this problem, linking users across two entirely different datasets.

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

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

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

We made a series of simple assumptions about human mobility. We broke time into discrete "bins" of a certain length, and then declared the number of checkins a user has at each location in time bin to be Poisson distributed according to a rate paramater λ unique to that time and place.

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

Table 1: Overview of datasets used in study. For FSQ-TWT and IG-TWT, number of locations refers to locations at a 4 decimal GPS granularity (position within roughly 10m).

This is a simple but reasonable assumption, and Poisson distributions are often used to model rare events (like checkins).

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

Figure 2 provides a visual illustration. On the left side of the image are two real world trajectories, denoted with a blue and turquoise line. The x axis shows space and the y axis shows time. The colored circles (red, green, gray and yellow) show times and places where the real world trajectories are sampled, with (for example) a geolocated photograph, phone call, or checkin. The challenge is that we only see the green, yellow, red, and gray trajectories in the middle of the image, and we must figure out the true association across datasets. In this example, red should go with green and gray with yellow. On the right side of the image, the concept of time bins are illustrated. We discretize time with varying sized time bins. The top uses one large time bin, essentially ignoring time, whereas the bottom breaks time into four sections, essentially saying two locations are only the same if the checkins occur near one another in time.

We evaluated this algorithm on multiple real-world datasets. Gathering the data in itself was a significant challenge, as each dataset needed to contain individuals with identities linked across two different data sources. Collecting information from one data source is enough of a challenge by itself, given unexpected and changing data formats, connectivity problems, rate limits, and more. Getting ground truth data across two datasets is thus more difficult, as two APIs need to be dealt with and user identities must be verified across the two.

We gathered three datasets:

- **Foursquare-Twitter** (FSQ-TWT): checkin data from the location-based social networking and review site Foursquare [3] and geotagged updates from the microblogging site Twitter [4]. This data was obtained in a prior work by other authors who allowed us to use their data [5]. We expect the behavior to be somewhat different across the two networks; Foursquare is primarily used to review restaurants, and Twitter is generally used.

- **Instagram-Twitter** (IG-TWT): Geolocated photographs from the image sharing site Insta-

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

[4] `twitter.com`

gram [5] and geotagged updates from the microblogging site Twitter. We first crawled Instagram, and then found users who had posted their Twitter usernames in their profiles. For each of these users, we used Twitter's API to crawl their public tweets. We expected this dataset to be the easiest to link, as there were high numbers of checkins on both sites for most users.

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

Statistics about these datasets is summarized in Table 1.

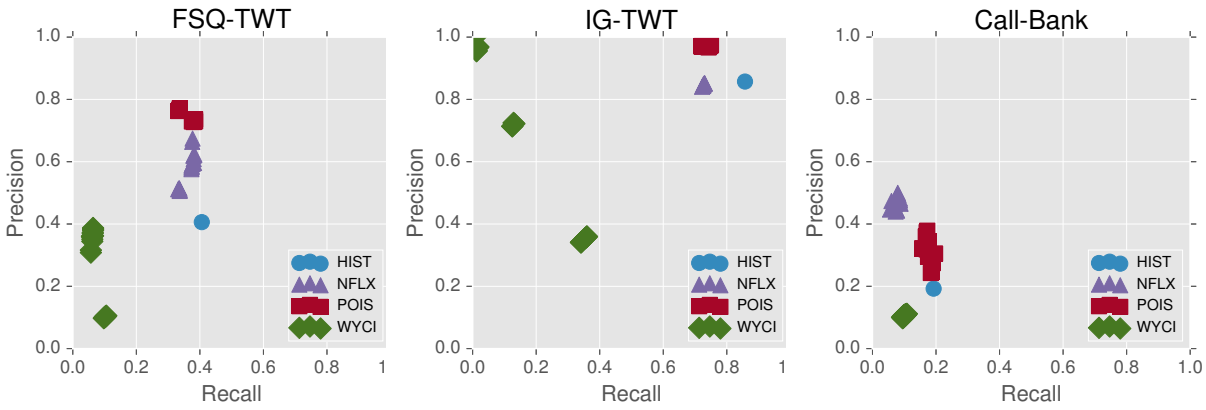


Figure 3: Precision and Recall plots for each dataset.

We now turn our attention to experimental performances of our algorithm. In Figure 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 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.

[5] `instagram.com`

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.

4.2.2 FindYou: A Personal Location Privacy Auditing Tool

5 Location Data, Demographics, and Bias

5.1 Related Work

5.2 Completed Work

5.2.1 I Dont Have a Photograph But You Can Have My Footprints

5.2.2 Scaling up the Census with Social Media

6 Proposal Topic I

As described in Chapter 5, an important challenge facing the computer science community is algorithmic bias. In recent years, an emerging body of work has focused on different mitigating techniques, such as automated discovery of bias, "de-biasing" existing algorithms, or theoretical analyses of differnet types of bias. De-biasing techniques are sure to incur a cost: the objective function of the algorithm is no longer as straightforward, and organizationally new infrastructure needs to be put into place for something that could hurt revenue. Understanding the key trade-offs between revenue and uncertain risk will be important to insure real-world adoption. Although there have been some good initial insights, the community has lacked strong data-driven analysis on this trade off.

I propose to fill this gap by applying proposed techniques to real-world problems through the use of an innovative dataset. Namely, I will look at the real-world problems of recommendation systems within a large social network. I will examine the trade off between recommendation accuracy, bias, and revenue.

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

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

The project will emerge in several stages.

1. Collection of instagram data (completed).
2. Labeling of instagram data with Face++ API (completed).
3. Initial analysis and descriptive statistics of dataset (in progress).
4. Full problem specificiation: algorithms, inputs, and objectives.

5. Apply de-biasing to algorithms and analyze impacts.

6. Create recommendations for algorithm designers.

7 Proposal Topic II

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

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

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

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

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

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

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

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

household with annual income below a certain threshold (TODO(verify, get numbers)). I will look for differences in the following behaviors:

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

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

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

8 Research plan

References

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

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

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

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

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