Big Location Data:
Balancing Profits, Promise, and Perils

Thesis Proposal

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

February 5, 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.

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

- Your Browsing Behavior for a Big Mac (maybe?)
- For Sale: Your Data. By: You (some portion?)
- Challenges of Keyword-Based Location Disclosure

4 Location Data and Anonymity

Linking Users Across Domains with Location Data
and
FindYou

5 Location Data, Demographics, and Bias

I Dont Have a Photograph But You Can Have My Footprints Under submission work on demographic labeling Current work on bias!

6 Proposal Topic I

7 Proposal Topic II

8 Research plan