Demographic Mobility

Research Document

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

May 5, 2016

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. However, due to difficulty in obtaining labels, little work has been done to understand the impact and importance of demographics on human mobility. In this thesis, we close this gap by pairing machine learning with large scale public social media data to label data and obtain new analyses of the impact of demographics on mobility.

Contents

1 Introduction

This part provides an overall introduction of your work, including related work of your proposal.

Ideas:

1. Real time polling: cheap, reliable, mostly accurate
2. Segregation
3. Bias
4. Fairness
5. Demographic inference from mobility
6. Semantics of locations
7. Location + Social Networks
8. Privacy
9. Uniqueness
10. Linking
11. Next place prediction
12. Inference and difficulties

Trying to cluster these... * Bias correction in polling * Human mobility and its uses... * Demographics * Fairness

1.1 Related work

This part talks about related work of your proposal.

2 Human mobility

Basics: Unique, social, sensitive, periodic

Location data can be described in two main ways: geographically or semantically (FIND A better WORD). Geographic data can be described by a latitude-longitude data on the globe. Semantic location data refers to an identifier commonly 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. Semantic location data 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.

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, except where noted.

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].

This section will contain info about location data. What are some features of human mobility? Privacy Social

3 Demographics

The word "demography" comes from the Greek words for "the people" and "measurement". Thus, demographics is the study of populations of human beings. In modern day usage, this typically involves

Some typical categorizations of demographies are:

- Race
- Socioeconomic status
- Gender
- Age
- Language

Demographics research is of interest for a variety of reasons.

Computational social science

Demographics is also deeply important in the world of advertising, the main driver of the online economy. Knowing the demographics of a site visitor means having a deeper insight into a reader's needs, desires, and hence their likelihood of purchasing something. Newspapers first made their money by providing "targeted" demographics, showing advertisers the addresses (and hence typically associated demographics) of readers. Now, through the use of computational techniques, ads can be targeted at users in much more fine-grained "buckets", the hope being that more fine-tuning results in higher click-through rates and thus higher revenues.

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

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

3.1 Segregation

U.S. Census has several exact definitions of housing segregation. The United States Census has

- Evenness
- Exposure
- Concentration
- Centralization
- Clustering

https://www.census.gov/hhes/www/housing/resseg/pdf/app_b.pdf

Recently, academic works have tried to use CDR data to understand segregation. [1] [6] [3] [2]

3.2 Inference

[8]

Differences in use on social networks.

4 Algorithmic Bias and Fairness

"Software is eating the world", Mark Andreesen famously said. As more parts of daily life become governed by software, the recommendations and algorithms within such products will have a larger impact on our society. Recently, concerns have been raised about algorithmic bias– the idea that the algorithms underlying our software may place disparate burdens or hardships on specific groups, particularly groups facing histories of discrimination or even legally protected classes. Perhaps concerningly, this bias can easily happen unintentionally or accidentally.

4.1 Evidence of Algorithmic Bias

Evidence or instances of algorithmic bias have been reported in the popular press. In 2012, the Wall Street Journal discovered that Staples was varying prices on their website, such that customers nearer to a competitor saw lower prices[3]. As customers in wealthier areas were more likely to be near a competitor, this had the effect of raising prices for lower income users. Prices are an extreme example, but bias can also extend to rankings or availability of products. More recently, Bloomberg reported on racial disparities in the ZIP codes where Amazon Prime Same Day delivery is available[4]. Reporters from the Washington Post showed that Uber wait times were longer in areas with higher concentrations of minorities [5].

[3] http://www.wsj.com/articles/SB10001424127887323777204578189391813881534

[4] http://www.bloomberg.com/graphics/2016-amazon-same-day/

[5] https://www.washingtonpost.com/news/wonk/wp/2016/03/10/uber-seems-to-offer-better-service-in-areas-with-more-white-people-that-raises-some-tough
http://www.nickdiakopoulos.com/projects/algorithmic-accountability-reporting/

Beyond the popular press, this topic has been studied in the academic community. In [7], authors found the same price changes mentioned in the Wall Street Journal article. Additionally, they found large price differences based on geographic location on several sites, especially for digital goods. Creating various personas (web-browsing histories of affluent or price-sensitive shoppers) led to changes in the ranking of goods on various sites, and prices differed based on the referring site of a user (such as a discount aggregator). They found no evidence for system-based (OS or browser) based changes.

Open data has allowed authors to look for bias in the methodology of law enforcement. In [4], researchers examined the impact of race on the Stop and Frisk policy of the New York Police Department. They were able to provide a simple procedure that police officers could take to greatly lower the number of stops unlikely to result in an arrest while keeping the number of illegal guns captured constant.

Authors have found some suggestions that companies might take care to not alter results when it comes to controversial topics [5].

4.2 Tools for Detecting and Correcting Algorithmic Bias

TODO FairTest Sunlight?

5 Surveying

6 Proposal Topic I

The first part of my thesis will be to obtain a demographic understanding of human mobility. In order to study demographic mobility, I must first obtain a dataset linking a user's demographics to their movements. I plan to do this by relying on the social network Instagram.

Instagram is a popular image-sharing social network, owned by Facebook. According to their press page [6], at the time of writing (April 2016), Instagram has over 400 million monthly active users, 75% of which are located outside the United States. Over 40 billion photos are shared on the site and 80 million photos are uploaded a day. Instagram first launched on smartphones, and as of writing there is no way to upload photos other than using a smartphone, making it very mobile-centric.

Instagram is invaluable for understanding demographic mobility for the following reasons. First, Instagram is rich in location data. Many users attach geographic information to their Instagram pictures, provided a sampled, active view of a their geographic location data. Some of these photos are also tagged with semantic location data. TODO: GIVE REAL NUMBERS

Second, Instagram's photos can hold the key to demographics. Recent advances in facial recognition have made it extremely easy and efficient to find faces in an image, and label the faces with age, race, and gender. It is therefore possible to label each public Instagram image with genders, ages, and races. By aggregating this information within a profile with many geotagged photos, we can potentially label the demographics of all of a users movements.

[6] `https://www.instagram.com/press/?hl=en`

After collecting this data, the next major hurdle will be to understand its representativeness and accuracy. To be more specific, we need to understand:

1. The **accuracy** of our algorithm labeling a user's age, gender, and race.
2. The **representativeness** of a user's movements within Instagram.
3. The **demographics** of Instagram users compared to the general population.

Each of these questions can be answered in the following manner:

1. **Accuracy of labels:** We can obtain an idea of accuracy by comparing our algorithm with profiles labeled by humans, either by research assistants or crowd-sourcing systems.
2. **Representativeness of mobility:** There are many works and datasets available on human mobility generated from a variety of behaviors, such as other social media, phone (CDR) data, taxi data, etc. We can compare our results with these other sources to find differences and similarities in mobility.
3. **Understand bias:** we can compare our Instagram results, or results from studies such as the Pew Internet Survey, with those from the U.S. Census or other surveys.

With an understanding of

- Create new metrics about the interactions of different demographics groups, relevant to sociologists.
- Run city planning surveys at fractions the cost of more expensive mail-in surveys.
- Better understand disparate impact of various services on different demographics.
- Use our understanding of demographics to assess the accuracy of algorithms labeling demographics in other contexts (e.g. on Twitter or other social media)

References

[1] Alexander Amini, Kevin Kung, Chaogui Kang, Stanislav Sobolevsky, and Carlo Ratti. The impact of social segregation on human mobility in developing and industrialized regions. *EPJ Data Science*, 3(1):1–20, 2014.

[2] Joshua Blumenstock, Ott Toomet, Rein Ahas, and Erki Saluveer. Neighborhood and network segregation: Ethnic homophily in a silently separate society. *Proc. NetMob*, 2015.

[3] Suma Desu, Lauren Alexander, and Marta Gonz'alez. Untangling the effects of residential segregation on individual mobility. *Proc. Netmob*, 2015.

[4] Sharad Goel, Justin M Rao, and Ravi Shroff. Precinct or prejudice? understanding racial disparities in new york city's stop-and-frisk policy. *Social Science Research Network*, 2015.

[5] Chloe Kliman-Silver, Aniko Hannak, David Lazer, Christo Wilson, and Alan Mislove. Location, location, location: The impact of geolocation on web search personalization. In *Proceedings of the 2015 ACM Conference on Internet Measurement Conference*, pages 121–127. ACM, 2015.

[6] Robert Manduca, Bradley Sturt, and Marta Gonz'alez. Mobile phone data as a means of stuyding activity space segregation at scale. *Proc. Netmob*, 2015.

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

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