The Ethics of Using Anonymized User Data

Mauro Hauptmann, Shazia Zaman, Daniel Ogier, Casey Burkett

[1] Southern Methodist University,
3300 University Blvd, Dallas, TX 75205

Abstract. Is it possible to identify Uber riders based on their pickup location in New York City, and what does this mean for privacy in our society as it becomes more cashless? Are there other privacy concerns apart from personal identification? In this document we look at publicly available data from Uber rides in New York City to identify riders as a proof of a loss of privacy a cashless society and look for other encroachments on privacy. We were able to reverse engineer Uber rider's identity from their pickup locations alone which shows that privacy is evaporating in our digital and increasingly cashless society. Additionally, we identified rider departure patterns from their homes.s Users of big data will have to exercise caution and a strong ethical focus to data they are given by users, knowing that their work could cause harm to others.

1 Introduction

As far back as 2000, before the onset of Big Data, anonymized data was shown to be useful in identifying individuals [4]. Now, we have more data than ever and more ways to link seemingly disparate information together. In this document we will look at the potential for privacy concerns in our increasingly cashless society. Cash provides truly anonymous transactions for buyers and sellers in a way many digital forms of payment don't. Cryptocurrencies can provide anonymity, but haven't yet gained wide acceptance, and can be broken to expose buyer and seller identities. In the following document we will look at a dataset of Uber (a popular and cashless ride sharing service) rides in New York City to determine if it is possible to identify riders from their pick up locations.

Recently, a lawsuit has alleged that Uber employees have used the company's data to track celebrities and other people of interest to them, and in this document we will demonstrate that special access to Uber's data isn't necessary to cause a violation of privacy for the company's riders. [1] There has been an increase in "ghost rides" (drivers charging fake rides to a compromised rider's account) [2], and the ability to accurately fake the locations could make it harder to identify fraud. Other researchers have been able to identify identities of DNA donors through the use of big data, which shows that using Uber data shouldn't be impossible. [3] The result of much previous work is that "anonymity doesn't ensure privacy" [5] and users of Big (and open) Data will have to exercise great caution using this data and consider the ethical ramifications of their work more than ever to prevent causing (or at least enabling) harm to unsuspecting users. Many users of any service have placed their trust with a service and the service provider has a duty to do the best it can to protect its users.

1.1 Dataset and Description

In this document, we analyze data from a dataset of Vehicle for Hire (VFH) rides in New York City (NYC). The data primarily focuses on Uber rides from April – September of 2014 and January – June of 2015. There are other VFH services included, but we are not analyzing that data because we cannot guarantee cashless payments in all cases, as we can with Uber. The data was obtained by FiveThrityEight (https://fivethirtyeight.com/) through Freedom of Information requests to the NYC Taxi and Limousine Commission (TLC). The files we used generally contained some combination fields for Date/Time, Latitude, Longitude, Dispatching base (which is a code from the TLC affiliated with the Uber pickup) and LocationID (an ID for Uber's pickup location) depending on the specific file. For detailed descriptions of the fields and data they contain, see the analysis section of this document.

1.2 Analysis Method Description

(Refine this as we solidify our work). For our analysis we used Python to read the dataset files into a MySQL database to compile all the pickup location records together then used the GoogleMaps API to match the location data to physical addresses. Once the addresses were obtained we removed commercial building pickups (hotels, for example) and used the White Pages API to cross reference the residential addresses to a resident(s). To see specific analysis implementation details, see the appendix.

2 Analysis of Uber Dataset

2.1 Feature Descriptions

The data from the csv files were loaded into several tables. A coordinates table that contained latitude and longitude as double values (ex. Latitude = 40.714224 and Longitude = -73.961452), and a pickup table that stores each pickup time, latitude, longitude, and base number (with latitude and longitude stored as in the coordinates table, and pickup time as a date and base as a string). This allows us to reference every ride to the coordinates and analyze the pickup locations and their frequencies as related to their general areas (using zipcodes). We then could pass the resulting addresses to the White Pages API that would return the resident's name.

2.2 Data Preparation and Cleaning

The steps to clean and prepare the data for analysis were:

1. Create a table called uber_Pickup with fields named Lat (for latitude), Lon (for longitude) and Base (for pickup base).

2. Load the 2014 data that followed the naming convention of uber-raw-data-(month)14.csv in to uber_Pickup
3. Create table coordinates with columns Processed (a flag indicated if the address has been retrieved
4. Select unique latitude and longitude information from uber_Pickup table and store it in the Coordinates table
5. Use the Google API to get addresses based on the latitude and longitude values (the API is found at https://developers.google.com/maps/documentation/geocoding/start). The next section shows an example of the formatted responses we collected from the API.
6. Create tables to store the collected information. These tables are: Neighborhood, containing neighborhood, sublocality, locality, county, state, county, lat, and lon; Street_address, containing street, city, state, zipcode, country, lat, lon, processed, and Station containing station, city, state, zipcode, country, lat, and lon.

2.3 Dataset Summary Statistics

As we converted the raw data from Uber into useable addresses we found 22,883 coordinates of latitude and longitude that we can reverse lookup into addresses and then into residents. Those addresses pointed us to 127 neighborhoods in NYC. We also found 1305 unique stations (meaning train or bus stations) that also show rider activity.

2.4 Pickup Location Analysis

Our dataset contained over 400,000 rows that had rides in 68,063 distinct addresses in the New York City area. Among the counties in the dataset, the borough of Brooklyn had the largest volumes of riders in the dataset.

The neighborhoods within Brooklyn that had the greatest amount of activity are Park Slope, Crown Heights, and Bedford-Stuyvesant (see graph below). Unfortunately, those areas are also very densely populated due to multifamily housing, such as apartments, that makes our identification difficult. It is likely that these areas had the most activity due to the high concentration of young due to the draw of a city center that is more attractive to younger people that are more frequent users of Uber.

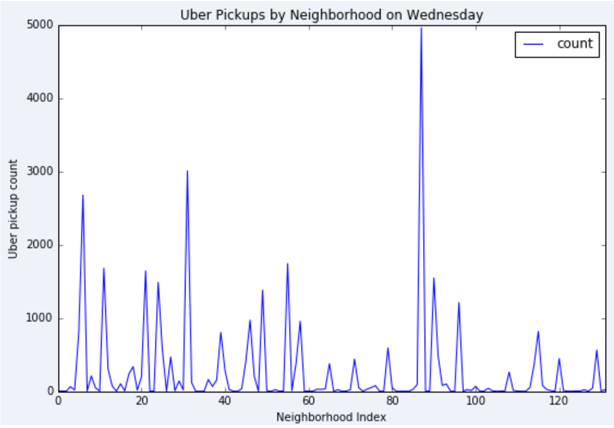

Fig. 1. Graph of Uber pickups on a Wednesday (to reduce tourists). The heaviest activity is in Brooklyn.

Looking at a lower level of detail among the dataset, we saw that the highest volume of rides tend to occur between noon and midnight indicating that Uber is used most in the afternoon and evening rather than the mornings. This could be due to common routes to business centers having decent mass transit options, but the unpredictability of evening trips elsewhere requiring a more flexible transportation option, like Uber.

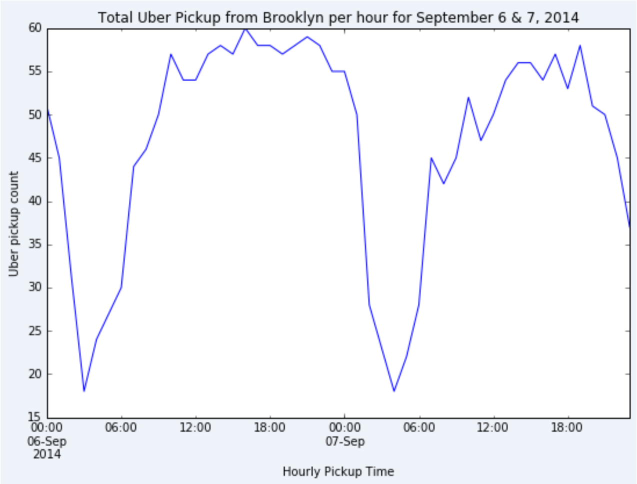

Fig. 2. Graph of pickup counts by hour on a Wednesday (to reduce tourists). Activity spikes around noon and tapers down to a low point around 6am.

2.5 Selection of Riders to Identify

We were limited in our ability to look up every single record we had by having only 1000 API calls that will turn our addresses into names. To choose our sample of 1000 addresses out of our entire dataset we looked at several factors that would help us choose identifiable addresses. New York City is very dense and has many commercial and multifamily buildings, that are not ideal marks for identification and thus we had to look around the metropolitan area to decide how to use our limited API calls wisely. We used a few different approaches to find more residential areas in the NYC area. We looked at home listings on Zillow.com to determine where more houses are listed for sale, versus areas where apartments or condominiums are listed. Looking at the population density in different zip codes in the area helped us focus on areas that would have the best chance of finding identifiable addresses, our assumption being that fewer apartments would result in a lower density. When using the API, any requests that returned an error didn't count against our 1000 calls so ultimately we selected over 3000 addresses to search.

From the API we collected JSON objects filled with arrays of other JSON objects using key value pairs that we parsed into our tables and the structure of the data returned made storing the data in our tables relatively simple.

2.6 Analysis of Identification and Rider Patterns

We were able to identify homeowners with all 1000 API calls that were allotted to us with varying amounts of success. Some of the results clearly identified a single resident home to match to our location data, while others identified many residents in a home. This means that not all 1000 addresses used were tied to a specific person, but the records we have for a single resident matched up with our validation against other data sources. To verify accuracy those sources we looked at appraisal district, census and tax data and a few free internet services that mimic the same process that the Whitepages API uses to find homeowner records via public government information. Initially, this was a fine result because our goal wasn't to identify every record successfully, but only to prove that this anonymized data could be used to violate a person's privacy. However we thought of another issue raised by the release and use of this data that was equally concerning.

For every record we had in our dataset, we looked at ride frequency at single family homes for the locations that had the most rides in our dataset. This showed us some consistent patterns in when people were leaving their homes. Some were in the morning, presumably for work, or at night. With this pattern in mind, it seems that it is possible to know when some Uber riders are consistently away from home, which is not information that anyone would like a stranger to know. This could be especially problematic if that stranger has ill intent combined with the knowledge of when a person is not home.

Below is a scatterplot that shows the rides from 3 addresses that had a high frequencies. You can see that the red triangles are clustered in the middle of the day and the green squares appear more at the end of the day (the top of the graph). This is not information that should be easily attainable for riders who didn't consent to the release of their data

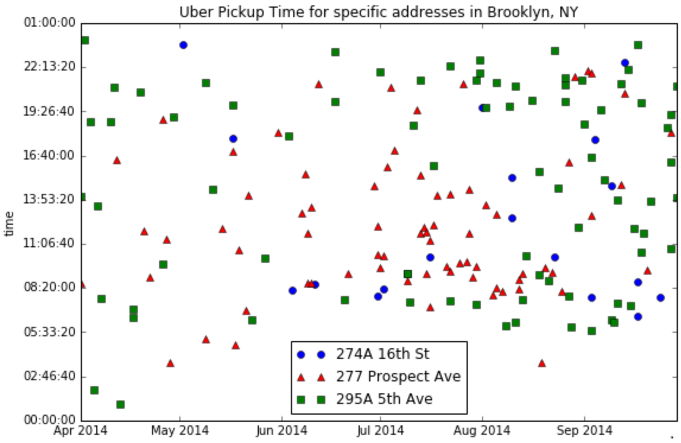

Fig.3 . Scatterplot of 3 addresses with high ride frequency. When looking at frequent riders, patterns emerge that reveal common departure times, which could be used to harm riders in some way.

3 Implications of Analysis

3.1 Societal Impact

As we have proven above using methods that are freely available, along with a no cost dataset, it is possible to violate a person's privacy with anonymous data that they didn't directly provide to us. When we consider the vast amounts of information we provide (both explicitly, and implicitly) to all manner of entities, it calls into question if it is possible to have true privacy in our society anymore. All our lives are lived in the digital as well as the physical world, and when data from both worlds comes together it is likely impossible to keep our personal information private. As more and more people move into fields of data analysis and big data, they will have the opportunity to violate people's privacy intentionally due to curiosity or malice, or ignorance, and negligence. These knowledge workers will be faced by a rising need for them to apply their skill ethically and act as a line of defense to the users who provided data in the first place. Users place great trust in many internet based services and it is the duty of the service to do their best to protect and earn the trust of their user base.

3.2 Conclusion

People in today's society are almost constantly surrendering data about themselves to a large array of entities that will have various intents for collecting it. The users should keep privacy in mind when choosing what services to use, but it would be very difficult and time consuming to monitor or withdraw from every digital service. Analysts of the data the users surrender also have an ethical responsibility to use the data for only its expressly stated purpose, and to disclose the data, and analysis in an ethical manner that protects the user and their privacy as much as possible. Simply anonymizing data is not always enough to protect users, as has been shown here.

2.5 Citations

[1]C. Lecher, "http://www.theverge.com," theVerge, 12 December 2016. [Online]. Available: http://www.theverge.com/2016/12/12/13920258/uber-employees-tracking-celebrities-security-lawsuit. [Accessed 24 January 2017]
[2] H. Taylor, "www.cnbc.com," CNBC, 19 January 2016. [Online]. Available: http://www.cnbc.com/2016/01/19/stolen-uber-accounts-worth-more-than-stolen-credit-cards.html. [Accessed 24 January 2017].
[3] DataFloq, "www.datafloq.com," DataFloq, [Online]. Available: https://datafloq.com/read/re-identifying-anonymous-people-with-big-data/228. [Accessed 24 January 2017].
[4] L. Sweeney, "Simple Demographics Often Identify People Uniquely," Carnegie Mellon University, Pittsburgh, 2000.
[5] S. Berinato, "www.hbr.org," Harvard Business Review, 09 February 2015. [Online]. Available: https://hbr.org/2015/02/theres-no-such-thing-as-anonymous-data. [Accessed 24 January 2017].

References

1. Baldonado, M., Chang, C.-C.K., Gravano, L., Paepcke, A.: The Stanford Digital Library Metadata Architecture. Int. J. Digit. Libr. 1 (1997) 108–121