FindYou: A Personal Location Privacy Auditing Tool

Chris Riederer
Columbia University
mani@cs.columbia.edu

Danny Echikson
Columbia University
dje2125@columbia.edu

Stephanie Huang
Columbia University
syh2115@columbia.edu

Augustin Chaintreau
Columbia University
augustin@cs.columbia.edu

ABSTRACT

The ubiquitous availability of location data to smartphone apps and online social networks has caused the collection of such information to grow at an unprecedented rate. However, the discriminative power and potential uses of this data collection is not always clear to the end user. In this work, 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. We hope that FindYou will increase awareness of the privacy issues surrounding the collection and use of location data, the potential problem of \digital red-lining", and also create valuable new datasets with the full informed consent of interested users.

1. INTRODUCTION

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 [15, 9, 14]. 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 [11]. In another, Mac users were shown more expensive hotels on a travel website [6].

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 [4, 13]. These tools, though very useful, are often not designed to inform non-technical users of the problems inherent in personalization.

In this work, 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, we will describe related work, the overall goals of our project, and the principles we focused on while designing it.

2. RELATED WORK

This project 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 [15, 9].

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2890546 .

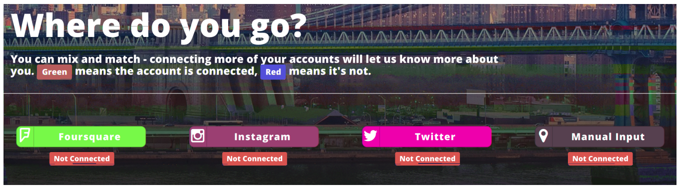
Other works have explored how users understand and value their location privacy [10]. In constrast to these works, we do not utilize user data as an object of study, or seek to understand user perceptions of location privacy. Rather, we wish to inform users about their location data and potential privacy 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 [3, 13], price discrimination [7, 11, 6], and ad targeting [5, 4, 12]. 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 [2, 8]. 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 [1]. Our project focuses on a specific subset of this larger goal, but offers a type of data that is not publicly widely available.

3. DESCRIPTION

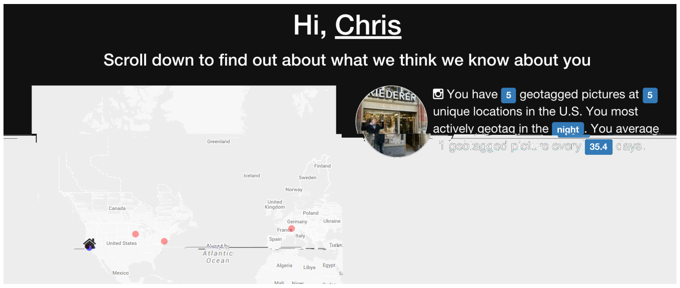
FindYou has two main goals: The first goal of our project is to inform users, regardless of technical skill, about what their location information can reveal. The second goal is to improve research on demographics and mobility by gathering a new dataset with the informed consent of interested users.



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 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

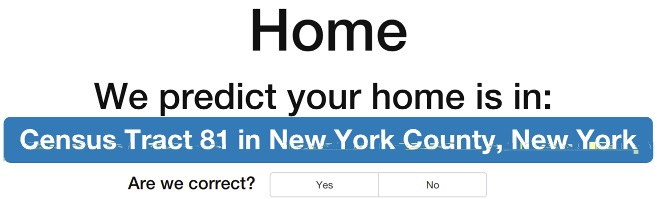
You have 5 geotagged pictures at 5 unique locations in the U.S. You most actively geotag in the night. You average 1 geotagged picture every 35.4 days.

Figure 2: After connecting their data, the user sees an overview of their locations and imported data.

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.

3.1 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.



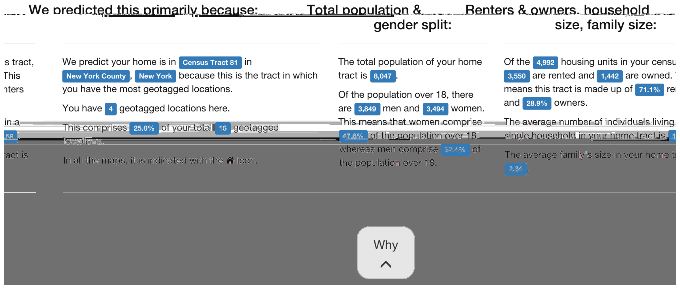
Home

We predict your home is in:

Census Tract 81 in New York County, New York

Are we correct? Yes | No

Figure 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 censu 3,550 are rented and 1,442 are owned. means this tract is made up of 71.1% re and 28.9% owners.

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

The average family's size in your home t 2.84.

Figure 4: For all predictions, we show additional details about how we made this guess.

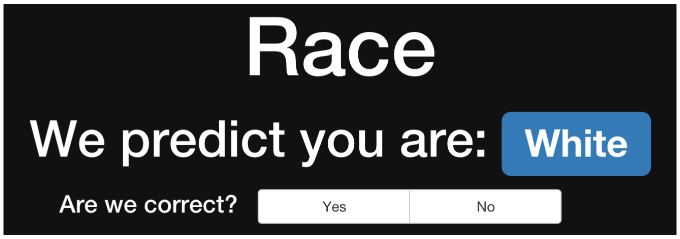
3.2 Design Decisions

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 webservices, as users must reveal their location to obtain any value from the service.

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, Twit-

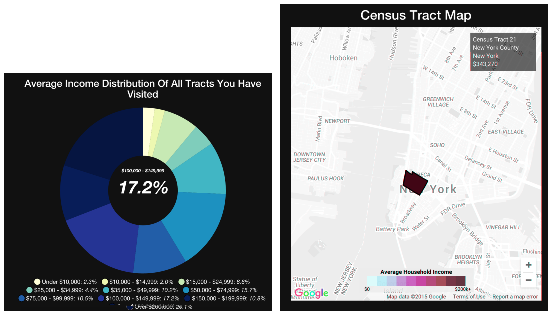


Race

We predict you are: White

Are we correct?

Figure 5: The site predicts several demographic attributes, one of which is race. The user has the option to tell us if we are correct.



(a)

(b)

Figure 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.

ter 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 the paper 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. 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.

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.

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`.

6. REFERENCES

[1] L. R. Goldberg, J. A. Johnson, H. W. Eber, R. Hogan, M. C. Ashton, C. R. Cloninger, and H. G. Gough. The international personality item pool and the future of public-domain personality measures. *Journal of Research in personality*, 40(1):84{96, 2006.

[2] Google. Google Location History. *https://www.google.com/maps/timeline*, Oct. 2015.

[3] A. Hannak, P. Sapiezynski, A. M. Kakhki, B. Krishnamurthy, D. Lazer, A. Mislove, and C. Wilson. 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, May 2013.

[4] M. Lecuyer, G. Ducoffe, F. Lan, A. Papancea, T. Petsios, R. Spahn, A. Chaintreau, and R. Geambasu. XRay: Enhancing the Web's Transparency with Differential Correlation . In *23rd USENIX Security Symposium (USENIX Security 14)*, San Diego, CA, 2014. USENIX Association.

[5] B. Liu, A. Sheth, U. Weinsberg, J. Chandrashekar, and R. Govindan. AdReveal: improving transparency into online targeted advertising. In *HotNets-XII: Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks*. ACM Request Permissions, Nov. 2013.

[6] D. Mattioli. On Orbitz, Mac Users Steered to Pricier Hotels. *online.wsj.com*, pages 1{6, Aug. 2012.

[7] J. Mikians, L. Gyarmati, V. Erramilli, and N. 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, Oct. 2012.

[8] Move-O-Scope. Move-O-Scope. *https://move-o-scope.halftone.co/*, Oct. 2015.

[9] C. Riederer, S. Zimmeck, C. Phanord, A. Chaintreau, and S. M. Bellovin. Si donSt have a photograph, but you can have my footprints."{revealing the demographics of location data. In *ACM Conference on Social Networks*, 2015.

[10] J. Staiano, N. Oliver, B. Lepri, R. de Oliveira, M. Caraviello, and N. Sebe. Money Walks: A Human-Centric Study on the Economics of Personal Mobile Data. *arXiv.org*, July 2014.

[11] J. Valentino-Devries, J. Singer-Vine, and A. Soltani. Websites Vary Prices, Deals Based on Users' Information. *online.wsj.com*, pages 1{6, Dec. 2012.

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

[13] X. Xing, W. Meng, D. Doozan, N. Feamster, W. Lee, and A. C. Snoeren. Exposing Inconsistent Web Search Results with Bobble. *Passive and Active Measurements Conference*, 2014.

[14] H. Zang and J. 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, Sept. 2011.

[15] Y. Zhong, N. J. Yuan, W. Zhong, F. Zhang, and X. 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.