“I don’t have a photograph, but you can have my footprints.”*

Chris Riederer* **Sebastian Zimmeck*** **Coralie Phanord**** **Augustin Chaintreau*** **Steve Bellovin***

*Columbia University

**Dartmouth University

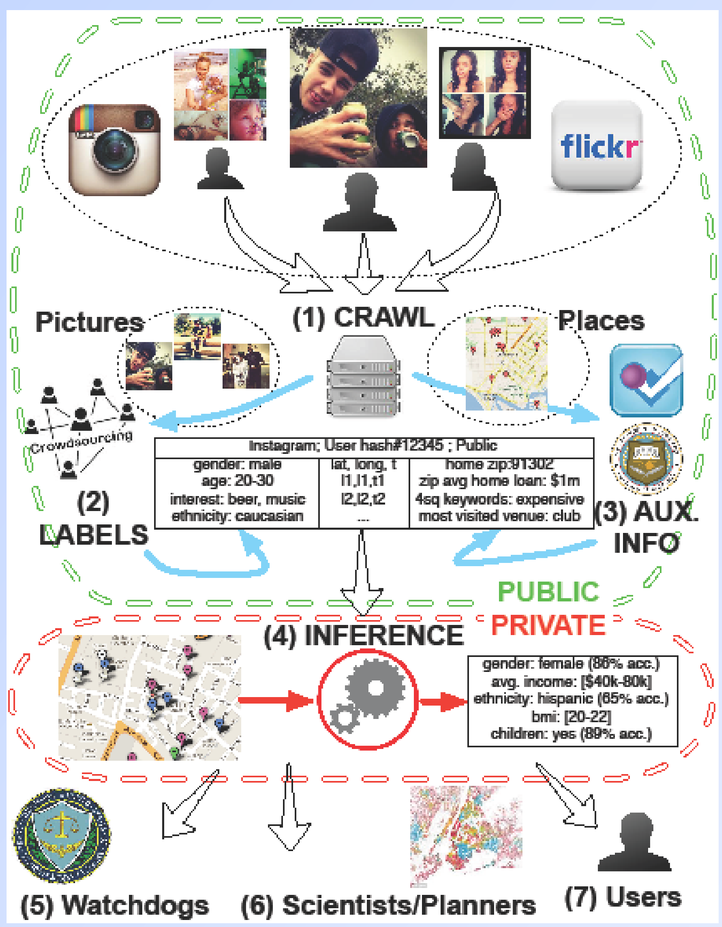
Problem

Location data is increasingly available and used to target advertising. While the identification risk of location information has been widely reported, its discriminative risk has received much less attention. We wish to determine what demographic traits can be inferred from users’ geographical footprints.

Methods

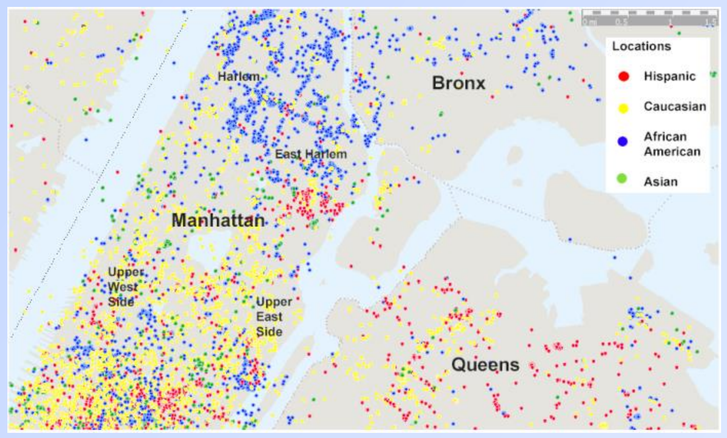
In order to address this problem we first obtain a dataset linking human mobility information to demographic attributes. To gather this dataset, we use the following methodology:

1. We obtain photographs and metadata from crawling photo-sharing services, such as Instagram or Flickr.
2. We pay crowd workers in an online marketplace, such as Amazon Mechanical Turk, to label the photographs with demographic information.
3. We obtain user location data from the metadata and augment it with auxiliary information, such as census data.
4. We combine all data for each user and use this information to train inference algorithms or observe mobility patterns.

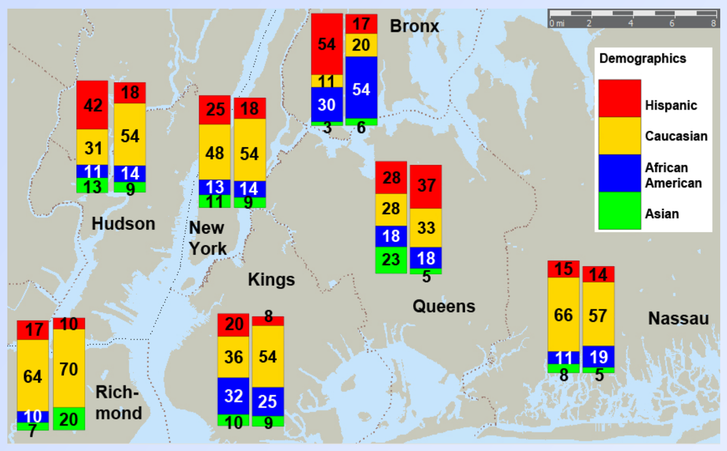


Analysis

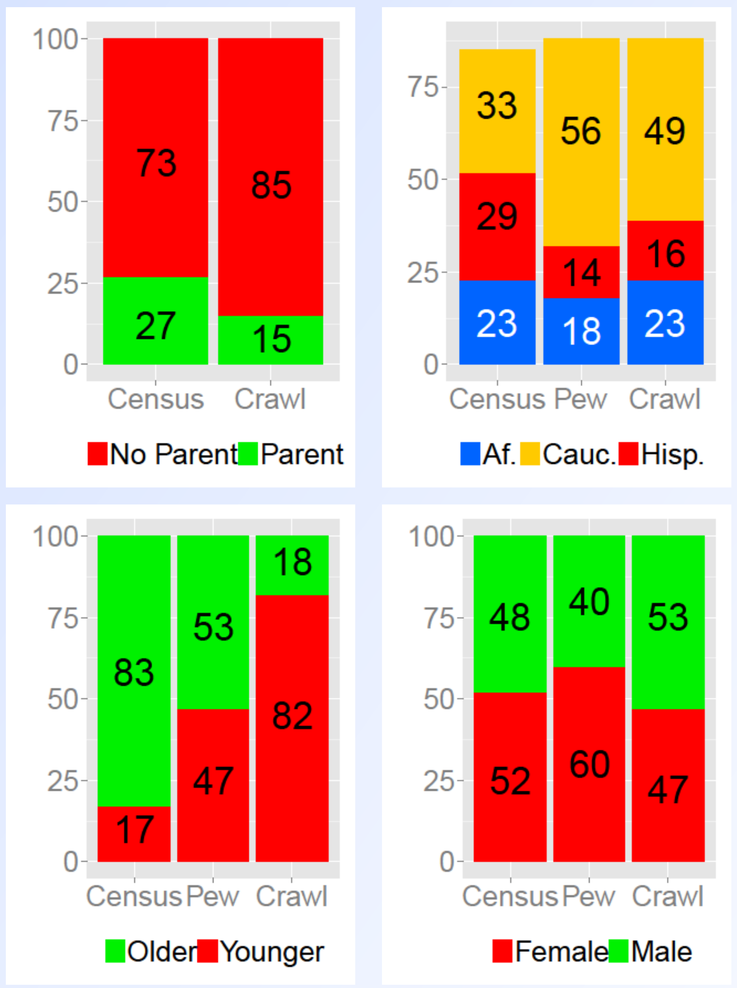
We compared our data with United States Census data. Our crawl appears to be consistent with much of the census information.



A dot map showing mobility patterns for several ethnic groups in New York City. Each dot represents a geotagged Instagram photo.



Comparison of ethnicity for seven New York state counties. (Right bars: our data; left bars: United States Census data)

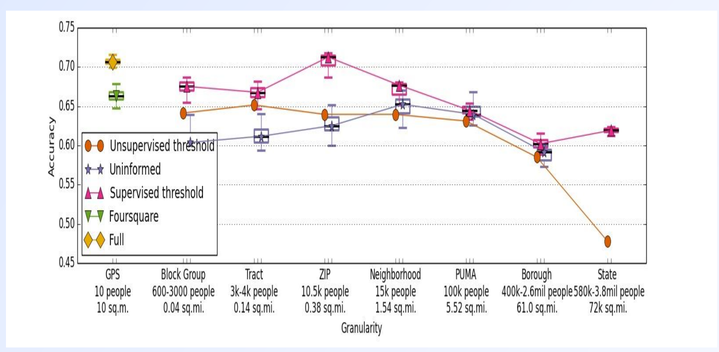


Comparison of United States Census data, Pew Research Center data, and our data (Crawl) for various demographic attributes.

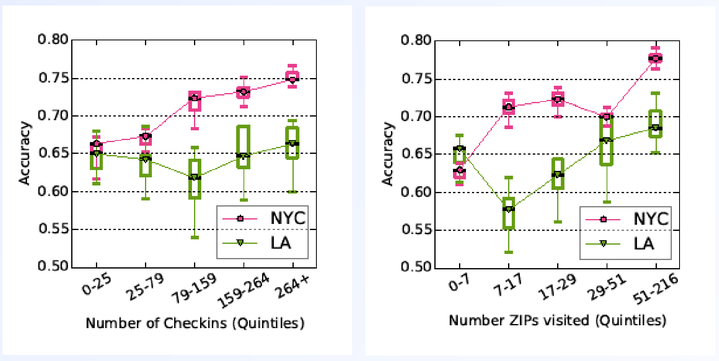
Inference

Our dataset allows us to explore several questions:

1. *Is inference possible?* We use both traditional machine learning techniques as well as simple decision rules.
2. *What is the impact of auxiliary information on inference?* We test both “informed” algorithms, which use features from the United States Census, and “uninformed” algorithms, which utilize only counts of visits to geographic areas.
3. *What is the impact of geographic granularity on inference?* Because our data is in precise latitude-longitude coordinates, we can decrease its granularity and observe the impact on accuracy.



The impact of location granularity on accuracy for several classes of algorithms.



Accuracy as a function of both amount of data (left) and diversity of data (right).

Future Work

There are many interesting directions to pursue in future work. First, we would like to examine more attributes to gain a deeper understanding of what discriminative information can be inferred about a user through their mobility. We would also like to understand the relationship between geography and inference. For example, our algorithms work much better for the New York area compared to the Los Angeles area. Finally, we wish to explore a theoretical basis for our results.

References

[*] Groucho Marx, A Night at the Opera

[1] E. Cho, S. A. Myers, and J. Leskovec. Friendship and Mobility: User movement in location-based social networks. In *KDD '11.*

[2] Y.-A. de Montjoye et al. Unique in the crowd: The privacy bounds of human mobility. Sci. Rep., 3, 2013.

[3] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences, 2013.

[4] H. Zang and J. Bolot. Anonymization of location data does not work: a large-scale measurement study. In MobiCom '11