Proceedings of the Ninth International AAAI Conference on Web and Social Media

"I Don't Have a Photograph, but You Can Have my Footprints"* — Revealing the Demographics of Location Data

Chris Riederer, Sebastian Zimmeck, Coralie Phanord, Augustin Chaintreau, Steven M. Bellovin
Computer Science Department, Columbia University, New York, NY
{mani,sebastian,augustin,smb}@cs.columbia.edu, Coralie.S.Phanord.16@dartmouth.edu

Abstract

High accuracy location data are routinely available to a plethora of mobile apps and web services. The availability of such data lead to a better general understanding of human mobility. However, as location data are usually not associated with demographic information, little work has been done to understand the differences in human mobility across demographics. In this study we begin to fill the void. In particular, we explore how the growing number of geotagged footprints that social network users create can reveal demographic attributes and how these footprints enable the understanding of mobility at a demographic level.

Our methodology gives rise to novel opportunities in the study of mobility. We leverage publicly available geotagged photographs from a popular photosharing network to build a dataset on demographic mobility patterns. Our analysis of this dataset not only reproduces previous results on mobility behavior at various geographical levels but further extends the existing picture: it allows for the refinement of mobility modeling from entire populations to specific demographic groups. Our analysis suggests the existence of regional variations in mobility and reveals statistically significant differences in mobility between genders and ethnicities.

1 Introduction

In recent years, the study of human mobility has flourished due to the proliferation of publicly available datasets. Many studies have delved into understanding how and why people move (González, Hidalgo, and Barabasi 2008; Noulas et al. 2011; Cho, Myers, and Leskovec 2011). However, much of this work has treated mobility patterns as homogeneous. In reality, different populations may have different movement behaviors. Previous studies were not able to explore any differences due to a lack of data. While some studies have shown that mobility is correlated to social status (Cheng et al. 2011) and community well-being (Lathia, Quercia, and Crowcroft 2012) measured at city and neighborhood levels, they are based on inferred attributes where the ground truth is not firmly established.

This paper explores how crowdsourcing and mobility patterns extracted from photosharing profiles enable the analysis of human mobility across demographics. Our exploration stands in contrast to limitations of previous studies as it brings together the following contributions:

- We show how photosharing network data can be leveraged to extract mobility patterns **introducing a new method for creating location datasets using publicly available resources**. (§2).
- Based on our created dataset we show that **mobility patterns extracted using our method can be partially validated** by comparing their essential characteristics to previous observations reported for Call Detail Records (CDRs). (§3.1).
- After making necessary adjustments to the labeled data from our dataset we show that the **ethnicity and gender distributions are similar to the corresponding distributions of the United States Census**. (§3.2).
- We analyze **ethnicity- and gender-specific mobility patterns on a demographic level and show statistically significant differences**. (§3.3).

Our study opens multiple avenues of research made possible by informative and publicly available location data. We provide some directions for future work alongside our results in our conclusion. (§4).

2 Methodology and Application

Public user profiles on photosharing networks often contain a significant amount of photos tagged with latitude-longitude locations. These data can be used to create comprehensive mobility profiles. Based on this insight we collected and labeled data using the following methodology.

Data Collection. We collected publicly available photo metadata from Instagram covering data for the years from 2011 through 2013. Although we used Instagram, it should be noted that this methodology can be also applied to other sites, such as Flickr or Facebook. As the metadata were publicly available our institution did not require Institutional Review Board (IRB) approval. We started a crawl from a root user (the founder of Instagram, on whose feed a large, diverse group of users comment) and followed further users subsequently through comments and likes. We skipped users with no geotagged photo in their first 45 photos. Our crawl

*G. Marx, *A Night at the Opera*.

Copyright © 2015, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

retrieved a total of 35,307,441 photo location data points belonging to 118,374 unique users.

User Labeling. After collecting the data, we labeled the ethnicity ($n = 1,015$) and gender ($n = 241$) of a subset of our users. To compare our results to previous studies (Isaacman and others 2010; 2011a; 2011b), we decided to label users from within the New York City (NY) and Los Angeles (LA) metropolitan areas. We selected users who had more than half of their *checkins* (that is, photos taken) within the region, which we believe sufficiently removed tourists. We then declared a user's home location to be the census tract in which he or she had the most checkins.

We then hired crowd workers on Amazon Mechanical Turk to annotate users' ethnicities and genders based on the users' photos. We asked the annotators to disregard accounts for businesses, celebrities, and others where they had doubt about the identity of the account owner and to make use of any tagged names to identify the account owners. Each user was labeled by two annotators. In cases of disagreement we asked a third annotator for an additional label to break the tie. If we did not obtain a majority agreement, we labeled the respective user ourselves.

For ethnicity labeling, we used the race and ethnicity categories of the United States Census 2010 (United States Census Bureau 2010): annotators categorized each user either as Hispanic or Latino (Hispanic), White alone (Caucasian), Black or African American alone (African American), one of the remaining census categories (Other), or "couldn't tell." For gender, we asked annotatorsN to categorize users as male, female, or "couldn't tell."

To measure the quality of agreement we used Krippendorff's α (Krippendorff 1980). Generally, α values above 0.8 are considered as good agreement, values above 0.67 as fair, and values below 0.67 as dubious (Manning, Raghavan, and Schütze 2008). We obtained α values of 0.74 for ethnicity in LA, 0.68 for ethnicity in NY, and 0.85 for gender in NY. We did not explore gender for LA.

3 Mobility and Demographic Patterns

We now present an analysis of mobility patterns on various population levels. Our dataset reveals mobility trends similar to CDRs (§3.1) and often represents the adjusted census population well (§3.2). In many cases we are able to detect differences in mobility patterns between ethnic groups and genders that can be plausibly explained by previous sociological findings (§3.3).

3.1 Mobility Patterns

In order to compare the mobility patterns of our dataset to those in the CDR dataset of (Isaacman and others 2010; 2011b) we only consider checkins for the years 2011 through 2013 each for the Spring months from March 15 to May 15 and for the Winter months from November 15 to January 31 (the LA and NY Spring and Winter subsets, respectively). Our data is more sparse: while the CDR dataset (Isaacman and others 2011b) has at least eight location points from call activity per day for the median user in LA and NY—and even 12 if text messages are added—the data in all of our subsets account for only one location point for the median user per day.

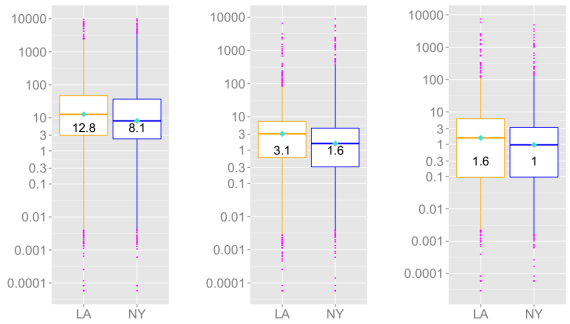

Figure 1: Daily ranges in miles. Boxes show the 25th, 50th, and 75th percentiles; whiskers the 2nd and 98th percentiles. The maximum range (Max. Mo.–Fr.) is the longest trip taken on a single day by a user for the entire Spring subset on a weekday, while the median range (Med. Mo.–Fr.) is the median distance (Isaacman and others 2010). The median range at night (Med. Night) represents the median distance a user has traveled on a day for the entire combined Spring and Fall subset from 7pm–7am (Isaacman and others 2011b). Our calculations do not consider any day where a user had a zero range, that is, had multiple checkins at the same location or a single checkin only.

An insightful metric to compare mobility patterns is the *daily range*, which is defined as the maximum straight line distance a phone has traveled in a single day (Isaacman and others 2010). Daily ranges can be characteristic for mobility because median daily ranges on weekdays represent a lower bound for a commute between home and work (Isaacman and others 2010). Figure 1 shows a subset of our results. Our ranges are generally smaller than those reported by (Isaacman and others 2010; 2011b). However, the general trends in both datasets are similar. Most importantly, people in LA have generally greater ranges than people in NY and both travel longer during the day than at night.

3.2 Demographic Patterns

While initial comparisons of our ethnicity and gender label distribution frequencies to the corresponding census frequencies reveal substantial differences, they can be explained by accounting for Internet and Instagram usage rates. For example, there are slightly more females than males (53% vs. 47%) in Bronx County according to the census (United States Census Bureau 2010), which is contrary to our observation that suggests substantially fewer females than males (38% vs. 62%). However, the usage rates of Internet (70% vs. 69%) (File 2013) and Instagram (16% vs. 10%) (Duggan and Brenner 2013) vary between females and males. In addition, while 86% of female account owners set their social network profile to private, only 74% of males do so (Madden 2012). Adjusting our observed frequencies for these differences leads to a distribution of females and males (50% vs. 50%) that is much closer to the census distribution.

Similarly to gender, we made adjustments to our labels

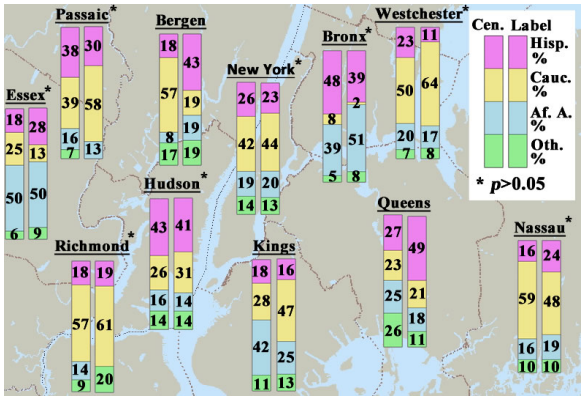

Figure 2: Detailed view of the multi-category ethnicity distributions for NY at the county level. Left bars show the census distributions (Cen.) and right bars the distributions from our labeling (Label).

for the varying percentages of Internet and Instagram usage among the different ethnicities as well. However, we still observed a substantial Hispanic underrepresentation, which was previously observed for Twitter in the American Southwest (Mislove et al. 2011). As ethnicity is not significant for setting a profile private (Lewis, Kaufman, and Christakis 2008), activity levels (posting pictures, etc.) are not lower for Hispanics (Statista 2012), and our annotation disagreements are not higher for the Hispanic label, this underrepresentation could be due to White Hispanics being perceived as Caucasian. This phenomenon has been observed before (McDonough and Brunsma 2013). So, we adjusted the observed frequencies by adding to the Hispanic labels a number of labels corresponding to the census percentage of White Hispanics and subtracting the same number from the Caucasian labels.

After adjusting our observed frequencies we perform chi square tests for goodness of fit comparing our ethnicity and gender distributions to the corresponding census distributions. We follow (Roscoe and Byars 1971) and require the average expected frequency for a chi square test with more than one degree of freedom to be at least two and for a test with one degree of freedom to be at least 7.5. To prevent skewing due to small sample sizes we use a Monte Carlo simulation with 2,000 replicates. Our results are promising. For example, Figure 2 shows that 8 out of 11 counties in the NY area had no evidence to reject the null hypothesis that the observed ethnicity distributions follow the corresponding census distributions (that is, $p > 0.05$).

3.3 Mobility Patterns by Demographic

By combining our methodologies from the previous two subsections we now show the differences in mobility patterns between ethnic groups and between males and females, respectively.

Daily Ranges. We calculate daily ranges for the different ethnic groups and genders based on our distributions of labeled users in LA and NY. More specifically, we obtain the same types of daily ranges as described earlier in Figure 1 (however, this time for all seasons of the year). Figure 3 shows some of our results. It is striking that Caucasians tend to have a much greater maximum daily range than the other ethnic groups. Indeed, a two sample Kolmogorov-Smirnov test reveals that the Caucasian range distribution differs significantly ($p < 0.05$) from the African American distribution and also from the Hispanic distribution. This result illustrates a more general finding: daily ranges of Caucasians often differ significantly from those of minorities. For 44% (8/18) of the comparisons of a Caucasian distribution to a minority distribution (three comparisons for maximum weekday, three for median weekday, three for median at night—each for LA and NY) the difference is significant at the 0.05 level. However, for the comparisons among minority distributions we only find 6% (1/18) to be significantly different from each other.

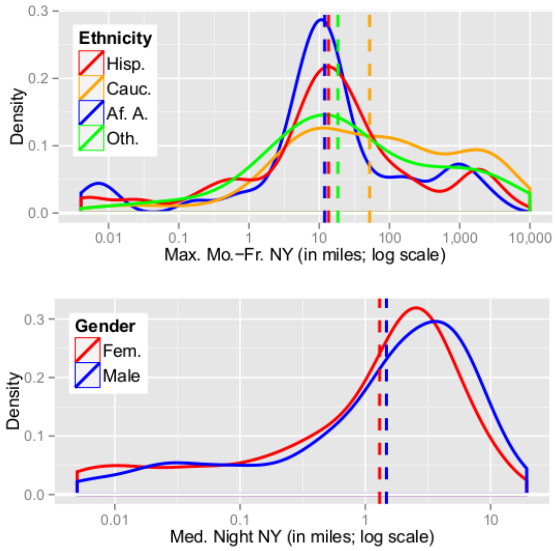

Figure 3: Daily ranges in miles. Top: density plot of the maximum daily ranges by ethnicity. Bottom: density plot of the median daily ranges at night by gender. We rounded daily ranges up to 0.005 miles and do not consider any day where a user had a checkin at only one location.

The differences in ranges by ethnicity can be most prominently observed in the comparisons of Caucasians to African Americans and Hispanics, respectively. However, it should be noted that *at night* all ethnicities have very similar ranges. This finding stands in contrast to the difference in daily ranges between males and females. In fact, the only statistically significant difference ($p < 0.05$) that we observed between males and females occurs for the median daily ranges *at night*. As shown in Figure 3, females tend to travel shorter distances at night than males. There are many possible explanations for this phenomenon. One reason could be that females travel less at night due to safety concerns (Badger 2014) and also avoid longer trips. In general, for both males and females—as well as for all ethnicities—we find that our observed daily ranges follow a (skewed) log normal distribution.

Home Ranges. In order to evaluate differences in mobil-

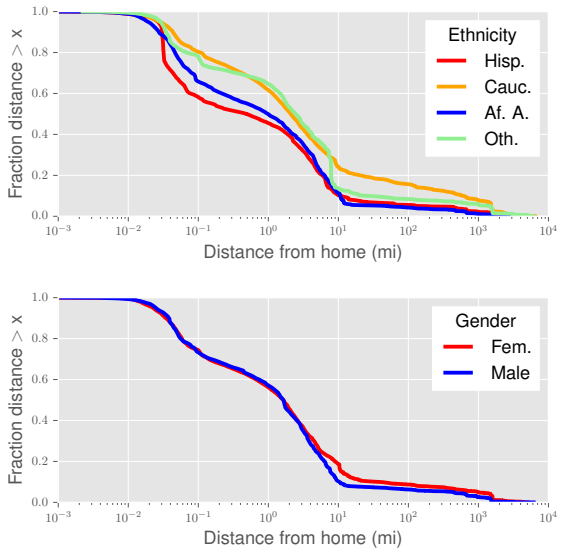

Figure 4: CCDFs of distance of checkin from home for NY users by ethnicity and gender.

ity with respect to someone's home location we complement the analysis of daily ranges with the evaluation of *home ranges*. A home range is a straight line distance between someone's home and another place to which the person traveled. Based on a user's home location, as specified in §2, we calculate the distance between the home and each checkin for the different ethnic groups and genders. Figure 4 shows the resulting CCDFs for the home ranges of the NY users.

Both graphs show a noticeable decrease around the 2,500 mile mark, which is the distance from NY to major hubs on the West Coast of the United States (e.g., LA with 2,475 miles). Males and females have similar home ranges, however, with females having more trips that are longer than about 8 miles. Possibly explanations might be that females take more vacations (Kelton 2013) and travel longer distances to work when they are employed full-time (Kwan 1999). These larger ranges are not inconsistent with the previous observation of shorter ranges for females at night as that result did not consider ranges during the day. The plot for ethnicity is in line with our previous observation that Caucasians travel farther from home than minorities.

4 Conclusion

The trove of geotagged pictures available through individual online profiles yields important insights for city planners and social scientists. It enables the extension of mobility analysis to demographics using shareable public datasets and reproducible results. Attributes such as age, occupation, or other lifestyle features could be extracted from users' photos and other mobility properties could be explored. Beyond this work we have started a systematic study of how mobility *alone* allows the inference of sensitive traits. Our work generalizes previous results on demographic inference from online activities. As mobility information becomes commonly available, we hope to make users better informed about the ramification of location disclosures.

References

Badger, E. 2014. http://wapo.st/1BmYO98.

Cheng, Z.; Caverlee, J.; Lee, K.; and Sui, D. 2011. Exploring millions of footprints in location sharing services.

Cho, E.; Myers, S. A.; and Leskovec, J. 2011. Friendship and mobility: user movement in location-based social networks. In *SIGKDD*, 1082–1090. ACM.

Duggan, M., and Brenner, J. 2013. The demographics of social media users - 2012. *Pew Research Center*.

File, T. 2013. Computer and internet use in the united states. http://www.census.gov/prod/2013pubs/p20-569.pdf.

González, M.; Hidalgo, C.; and Barabasi, A.-L. 2008. Understanding individual human mobility patterns. *Nature*.

Isaacman, S., et al. 2010. A tale of two cities. In *HotMobile '10*. ACM Request Permissions.

Isaacman, S., et al. 2011a. Identifying important places in people's lives from cellular network data. *Pervasive Computing*.

Isaacman, S., et al. 2011b. Ranges of human mobility in Los Angeles and New York. In *Pervasive Computing and Communications Workshops*.

Kelton. 2013. 4th annual springhill suites annual travel survey. http://bit.ly/1N2DdsX.

Krippendorff, K. 1980. *Content analysis: An introduction to its methodology.* Beverly Hills, CA, USA: SAGE.

Kwan, M.-P. 1999. Gender, the home-work link, and space-time patterns of nonemployment activities. *Economic Geography*.

Lathia, N.; Quercia, D.; and Crowcroft, J. 2012. The hidden image of the city: sensing community well-being from urban mobility. In *Pervasive computing*. Springer. 91–98.

Lewis, K.; Kaufman, J.; and Christakis, N. 2008. The taste for privacy: An analysis of college student privacy settings in an online social network. *J. Computer-Mediated Communication* 14(1).

Madden, M. 2012. Privacy management on social media sites. *Pew Research Center*.

Manning, C. D.; Raghavan, P.; and Schütze, H. 2008. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.

McDonough, S., and Brunsma, D. L. 2013. Navigating the color complex: How multiracial individuals narrate the elements of appearance and dynamics of color in twenty-first-century america. In Hall, R. E., ed., *The Melanin Millennium*. Dordrecht: Springer.

Mislove, A.; Lehmann, S.; Ahn, Y.-Y.; Onnela, J.-P.; and Rosenquist, J. N. 2011. Understanding the Demographics of Twitter Users. In *ICWSM'11*.

Noulas, A.; Scellato, S.; Mascolo, C.; and Pontil, M. 2011. An empirical study of geographic user activity patterns in foursquare.

Roscoe, J. T., and Byars, J. A. 1971. An Investigation of the Restraints with Respect to Sample Size Commonly Imposed on the Use of the Chi-Square Statistic. *Journal of the American Statistical Association* 66(336):755–759.

Statista. 2012. Social networking time per user in the united states in july 2012, by ethnicity (in hours and minutes). http://bit.ly/1GHojqn.

United States Census Bureau. 2010. 2010 census. http://factfinder2.census.gov/faces/nav/jsf/pages/index.xhtml.