"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

Location data are routinely available to a plethora of mobile apps and third party web services. The resulting datasets are increasingly available to advertisers for targeting and also requested by governmental agencies for law enforcement purposes. While the re-identification risk of such data has been widely reported, the *discriminative* power of mobility has received much less attention. In this study we fill this void with an open and reproducible method. We explore how the growing number of geotagged footprints left behind by social network users in photosharing services can give rise to inferring demographic information from mobility patterns. Chiefly among those, we provide the first detailed analysis of *ethnic* mobility patterns in two metropolitan areas. This analysis allows us to examine questions pertaining to spatial segregation and the extent to which ethnicity can be inferred using *only* location data. Our results reveal that even a few location records at a coarse grain can be sufficient for simple algorithms to draw an accurate inference. Our method generalizes to other features, such as gender, offering for the first time a general approach to evaluate discriminative risks associated with location-enabled personalization.

Categories and Subject Descriptors

K.4.1 [**COMPUTERS AND SOCIETY**]: Public Policy Issues—*Privacy*

Keywords

Location Data; Machine Learning; Privacy; Segregation; Social Networks

1. INTRODUCTION

Human mobility is intimately intertwined with highly personal behaviors and characteristics. As Justice Sotomayor of the United States Supreme Court stated, "disclosed in [GPS] data ... [are] trips the indisputably private nature of which takes little imagination to conjure: trips to the psychiatrist, the plastic surgeon, the abortion clinic, the AIDS treatment center, the strip club, the criminal defense attorney, the by-the-hour motel, the union meeting, the mosque, synagogue or church, the gay bar and on and on [47]." For that reason, previous studies of mobility centered on the risk of either re-identification in sensitive anonymized location datasets or on protecting visits to private locations [9, 16].

However, the re-identification risk based on individual locations is not the only threat. Many users are producing a series of footprints, which might be innocuous individually, however, taken together can create a sparse yet informative view allowing inferences from their whereabouts. The benefits of revealing locations are obvious: location data can be used for personalizing recommendations [39] and displaying more relevant advertising [28] in order to finance free online services. However, the downsides are more difficult to assess. While an individual data point may create no privacy risk, an aggregated dataset might enable inferences beyond a user's expectation.

In this paper we explore the discriminative power of location data. Solely based on mobility patterns, which we extracted from photosharing network profiles, we infer users' ethnicities and gender both on a demographic and an individual level. As we discuss in §2, this exploration stands in contrast to limitations of previous studies as our paper brings together the following contributions:

- We show how photosharing network data can be leveraged to extract mobility patterns using a new method for creating location datasets from publicly available resources. Our method combines the use of online social networks and crowdsourcing platforms. It has the advantage that it generally enables *anyone* to study human mobility and does not mandate access to Call Detail Records (CDRs) or other proprietary datasets. (§3).
- To assess the quality of the created datasets we show that mobility patterns extracted from photosharing networks are comparable in terms of their essential characteristics to those previously observed and reported for CDRs. For the first time, we extend the analysis of mobility patterns to *ethnic groups*. We show how comparisons lead to statistically significant differences that are meaningful for assessing residential and peripatetic segregation. (§4).
- Finally, we demonstrate the discriminative power of location data on an *individual* level. Our analysis confirms for the first time that location data alone suffices to predict an individual's ethnicity, even with relatively simple frequency-based algorithms. Moreover, this inference is robust: a small amount of location records at a coarse grain allows for an inference competitive with more sophisticated methods despite of data sparsity and noise. (§5).

*Groucho Marx, *A Night at the Opera*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
COSN'15, November 2–3, 2015, Palo Alto, California, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3951-3/15/11 ...$15.00.
DOI: http://dx.doi.org/10.1145/2817946.2817968.

2. RELATED WORK

Our study complements works on human mobility patterns and attribute inference in multiple ways.

First, the use of location data relates our study to previous inquiries into human mobility [7, 14, 36]. In particular, we aggregate location data into mobility patterns and compare our patterns to those published in earlier studies [3, 20, 21] for validation, but furthermore we analyze those patterns both at an individual level and aggregated in multiple demographic groups, including, for the first time, from the perspective of ethnicity. This analysis complements previous studies which have shown that mobility is correlated to social status [6] and community well-being [25] measured at city and neighborhood levels. While some studies already demonstrated that mobility traces can uniquely identify individuals [9, 44], the inference of individuals' demographic attributes from location data, that is, the *discriminative* power of location data, remained unexplored. We make inferences beyond trip purpose identification [11], activity type prediction [27, 29], and identification of location types [19].

Previous studies aimed to infer the ethnicities, gender, and other attributes of online users. Often they leveraged linguistic features, such as Facebook or Twitter user names, stated first and last names [5, 35], or Tweet content [39, 40]. Those studies demonstrated an underrepresentation of females and minorities online [35]; a finding which we extend and confirm using photosharing services. Mobility data from mobile phones were used to predict personality traits [10], age [4], and gender [43], but, in addition to relying on proprietary data, all of these studies solely analyzed call patterns or social network properties as opposed to locations. In contrast, we attempt to infer attributes using *only* location data, making our work more broadly applicable to any technology that can collect mobility information, such as GPS, Wi-Fi, or mobile apps. We additionally examine whether predictions become more accurate with more data, similar to [1], and how the granularity of data impacts prediction accuracy.

More generally, our analysis fits into the category of works on extracting information from social networks, such as [8]. Probably, the closest work is [50], which also aims to infer meaning from locations, however, is not concerned with ethnicity. We obtain our data from profiles of the photosharing service Instagram, and our analysis is enhanced with auxiliary information from the geo-social search service Foursquare and the United States Census 2010 [46] (Census). To our knowledge this is the first study demonstrating that it is possible to extract from social networks mobility patterns that are enriched with ethnic or gender information at an individual level. It should be noted in particular that all aforementioned studies of mobile data rely on proprietary data, primarily CDRs, that are only available with the consent of the data owner (e.g., [9, 25]). In contrast, our methodology is principally reproducible by anyone at a small cost, and our data will be made available shortly after publication. Our study provides a contribution to overcome the lack of publicly available mobility datasets and serves as a validator for their patterns.

3. METHODOLOGY AND APPLICATION

User profiles on photosharing networks often contain a significant amount of photos tagged with latitude-longitude GPS locations. Over time the accumulated location data can build up to comprehensive mobility profiles. Based on this insight and given that many user profiles on photosharing networks are publicly accessible we now introduce a methodology and its application to construct mobility datasets from readily available data. An overview of our methodology is shown in Figure 1.

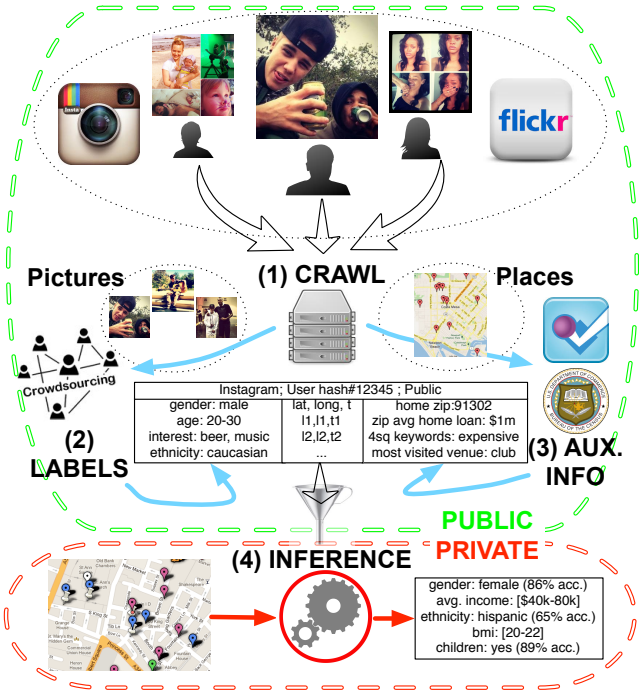


Figure 1: Methodology overview. A mobility dataset can be built in the following steps: (1) Public user profiles of a photosharing service are crawled and photo metadata are extracted into a database (Data Collection). (2) Corresponding photos are labeled (with labels for ethnicity, gender, etc.) by crowd workers in an online labor marketplace (User Labeling). (3) The dataset is further enhanced with auxiliary data, e.g., with the information that a certain location is close to a restaurant (Adding Auxiliary Information). (4) The dataset can then be used to analyze attributes on various demographic levels or train and test classifiers for individual inferences.

Data Collection.

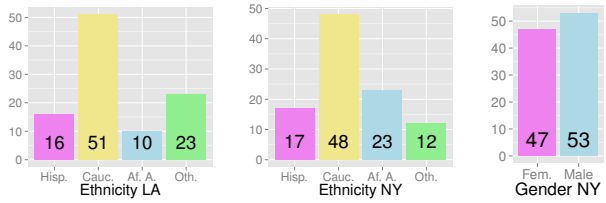
Applying this methodology, we collected publicly available photo metadata from Instagram covering data for the years from 2011 through 2013. This data collection and use was exempt from user informed consent under our institution's IRB rules since (1) we only collected publicly available online metadata, (2) after we used the metadata and the users were labeled, any identifying information, such as usernames, were removed, and (3) we only kept track of users' identities separately and for one single purpose (ensuring that the data we collected still belongs to a public Instagram profile). We started our crawl from a root user (the founder of Instagram, on whose feed a large and 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 retrieved a total of 35,307,441 photo location points belonging to 118,374 unique users.

User Labeling.

To match previous studies [19, 20, 21] that leveraged ZIP codes of CDR billing addresses from the Los Angeles (LA) and New York City (NY) metropolitan areas we randomly chose users from those areas as well. A user's home is the ZIP code where he or she had the most checkins (that is, photos taken). Note that this mitigates the content produced by tourists and other occasional visitors to LA and NY unless those have no other Instagram activity. A com-

bination of workers on Amazon Mechanical Turk (MTurk) and undergraduate students were asked to annotate users' ethnicities and gender based on the users' photos. However, in order to ensure that user pictures on Instagram profiles are sufficient to make a conclusive determination of users' ethnicities and genders we ran a preliminary experiment by selecting 200 profiles at random (excluding celebrities and business accounts) and having each labeled independently by two undergraduate students. We observed a strong agreement on gender (98%). The errors corresponded to a family profile belonging to multiple people and profiles with one picture.

For ethnicity labeling we leveraged Census categories. We asked the student annotators to categorize each user either as Hispanic or Latino (Hispanic), White alone (Caucasian), Black or African American alone (African American), or Other (combining all remaining Census categories, including Asian). Merging all remaining Census fields in the last category limits our detail view, although we would otherwise have some annotations being quite rare. Just as in the Census, our Hispanic category includes Hispanics and Latinos of any race, while the remaining categories do not include any Hispanics or Latinos. We found that our profiles are diverse: 45% Caucasian, 21% Hispanic, 15% African American, and 19% Other. The students' labels matched 87% of the time and when evaluated as a binary classification task (Caucasian vs. all other categories) the agreement reached 94%. It should be noted that the two labeling students were of different gender and ethnicity themselves. In conclusion, despite sparse data and ethnicity spanning a continuous spectrum, we found that labels are surprisingly predictable and consistent across annotators. As studies confirmed that 91% of teens post a photo of themselves on social networks [31] and that 46.6% of photos are either selfies or show the user posing with other friends [17] there is also evidence in many cases that it is actually the account owner who is shown in the pictures.



	Ethnicity LA	*Ethnicity NY*	*Gender NY*
Users (n)	427	588	241
K.'s α Multi.	**0.74**	**0.68**	-
K.'s α Binary	**0.78**	**0.74**	**0.85**

Figure 2: Annotations for LA and NY. Top: percentages of user labels for the different categories. Bottom: absolute numbers of labeled users and annotation agreement results.

To scale our annotation, we asked MTurk annotators to label a larger number of profiles for the same metropolitan areas using the same label categories. For consistency, we did not reuse the profiles used for the preliminary experiment described above. Each profile was labeled by two MTurk annotators. In cases of disagreement between the MTurk annotators we asked one of our undergraduate annotators for an additional label to break the tie or assign a label from a different third category. We decided to use a tiered annotation mechanism with the undergraduate annotator making the final decision in case of disagreements as unsupervised crowd workers on MTurk or similar platforms tend to be less attentive than physically available workers [37], who also have the possibility to ask clarifying questions. We were also careful to not drop any labels to avoid the introduction of a systematic annotation bias. Over two days 117 MTurk annotators participated in our task resulting in 1,015 properly labeled users with the labels shown in Figure 2. On the first day the annotators were compensated $0.10 per annotation and on the second day $0.05. The undergraduate annotator was compensated the regular stipend at our institution.

In order to measure the quality of agreement among the annotators we made use of Krippendorff's α [23]. Generally, values above 0.8 are considered as good agreement, values between 0.67 and 0.8 as fair agreement, and values below 0.67 as dubious [32]. Figure 2 shows that we obtained fair and good agreement and, thus, reliable ground truth for both our ethnicity and gender classifications.

Adding Auxiliary Information.

We collected auxiliary information from two sources. First, for the comparative analysis of demographic patterns with our data in §4.2 we used data from the Census [46] to associate geographic regions with gender and ethnicity distributions. Throughout the study we use Census-defined geographic granularities, ranging from block groups of 600-3k people to neighborhood tabulation areas (NTAs; 15k people), public use microdata areas (PUMAs; 100k people), and counties with populations of up to 2.6 million. We adjusted the distributions by ethnicity- and gender-specific Internet [13, 30] and Instagram [12] usage numbers. As explained in §4.2 we also took into account that Caucasian Hispanics are often perceived as Caucasian alone [34]. Second, for each checkin we obtained Foursquare information on the ten closest venues. We then used Foursquare's average venue popularities and venue categories as features for our inference algorithms (§5) since those features could provide an estimate of the types of places a user would visit.

4. MOBILITY-DEMOGRAPHICS

We now present a mobility pattern analysis for various population levels. Our dataset reveals mobility trends similar to those of CDRs (§4.1) and generally represents the adjusted Census population well (§4.2). 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 (§4.3), and we are also able to detect segregation among ethnic groups (§4.4).

4.1 Mobility Patterns

In order to compare the mobility patterns of our dataset to those in the CDR dataset of [20, 21] 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). Table 1 shows the distribution of the data in our subsets compared to those in the CDR dataset [20]. The mobility traces from our subsets are much more sparse. Most notably, while the CDR dataset 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.

Another insightful metric for comparing mobility patterns is the *daily range*, defined as the maximum straight line distance a phone has traveled in a single day [21]. Daily ranges are characteristic for mobility because, for example, median daily ranges on weekdays represent a lower bound for a commute between home and work locations [21]. Figure 3 shows a subset of our results. Our ranges are generally smaller than those reported by [20, 21]. However, the general trends in both datasets are similar. Most importantly, people in LA have generally greater ranges than people in NY. Also, in both areas people tend to travel longer during the day than at night. However, there are also differences: according to our data New Yorkers in the 98th percentiles travel farther than Angelinos.

	Spring		*Winter*	
Statistic	LA	NY	LA	NY
Total Checkins (Total CDRs)	135,503 (74M)	109,506 (62M)	118,446 (247M)	98,286 (161M)
Min. Loc./Day	1	1	1	1
1st Qu. Loc./Day	1	1	1	1
Med. Loc./Day (Med. Calls/Day) (Med. Texts/Day)	**1** **(9)** -	**1** **(10)** -	**1** **(8)** **(4)**	**1** **(9)** **(3)**
Mean Loc./Day	1.97	2.12	1.96	2.1
3rd Qu. Loc./Day	2	2	2	2
Max Loc./Day	73	62	98	69

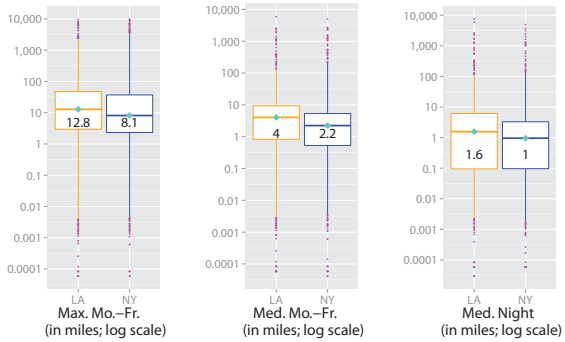
Table 1: Statistics of our LA and NY subsets compared to the CDR dataset in [20] (where available, in parentheses). Our calculations do not consider any day where a user had no checkins.

4.2 Demographic Patterns

As our LA and NY subsets are annotated with ethnicity and gender labels (§3) we are able to compare the resulting demographic distributions to the respective Census distributions. However, initial comparisons reveal substantial differences. For example, according to the Census there are more females than males (53% vs. 47%) living in Kings County [46] while our observed label frequencies suggest that there should be substantially fewer (43% vs. 57%). This result is even more surprising as the gender-specific usage rates of Internet (70% vs. 69%) [13] and Instagram (16% vs. 10%) [12] should further increase the percentage of females beyond the Census. However, while 86% of female social network account owners set their profile to private, only 74% of males do so [30]. Adjusting the Census distribution for this difference (as well as for gender-specific Internet and Instagram usage rates) leads to a distribution of females and males (49% vs. 51%) much closer to the distribution we observed for our labels.

Similarly to gender, we make adjustments to the Census distributions for the varying percentages of Internet and Instagram usage rates among different ethnicities as well. However, even then we still observed a substantial Hispanic underrepresentation, which was also observed for the southwest of the United States by [35]. We found this phenomenon difficult to assess, specifically, as ethnicity is not significant for setting a profile private [26], activity levels (posting pictures, etc.) are not lower for Hispanics [45], and our annotation disagreements are not higher when the Hispanic label is involved. However, we believe that the reason for the underrepresentation is the perception of Caucasian Hispanics as Caucasian alone. In a study, six of seven Caucasian Hispanics reported that others see them as Caucasian alone [34]. Therefore, we believe that most Caucasian Hispanics were actually labeled as Caucasian (i.e., our annotators agreed on an incorrect classification). Thus, we adjusted the observed label frequencies by adding to the Hispanic labels a number of labels corresponding to the Census percentage of Caucasian Hispanics and subtracting the same number from the Caucasian labels.

We perform chi square tests for goodness of fit comparing the gender and ethnicity distributions of our labels to the corresponding Census distributions for different levels of granularity. In most cases we obtain a value of $p > 0.05$ and find no evidence to reject the null hypothesis that the observed gender and ethnicity distributions follow the corresponding Census distributions. For example, as shown in Figure 4, for eight out of 11 counties in the NY area our tests resulted in $p > 0.05$ providing no evidence that our multi-category ethnicity distributions deviate significantly from the Census distributions. However, there are also cases with differences. It is no surprise that this is true for the state level as our distributions only cover users from the LA and NY metropolitan areas. However, overall we believe our results suggest that geotag data often replicate demographic trends faithfully.



	Max. Mo.–Fr.		*Med. Mo.–Fr.*		*Med. Night*	
%	LA	NY	LA	NY	LA	NY
98	2,471.7 (2,467)	3,625.6 (2,455)	133 (32)	209.9 (29)	117.4 (23.1)	129.9 (19.4)
75	47.5 (130)	37 (111)	9.3 (10)	5.3 (8.2)	6.1 (8)	3.3 (5.6)
50	**12.8** (36)	**8.1** (27)	**4** (5)	**2.2** (3.8)	**1.6** (4)	**1** (2.6)
25	3 (17)	2.3 (12)	0.8 (2)	0.5 (1.3)	0.1 (1.4)	0.1 (0.7)
02	ϵ (1.6)	ϵ (1.3)	ϵ (0)	ϵ (0)	ϵ (0)	ϵ (0)

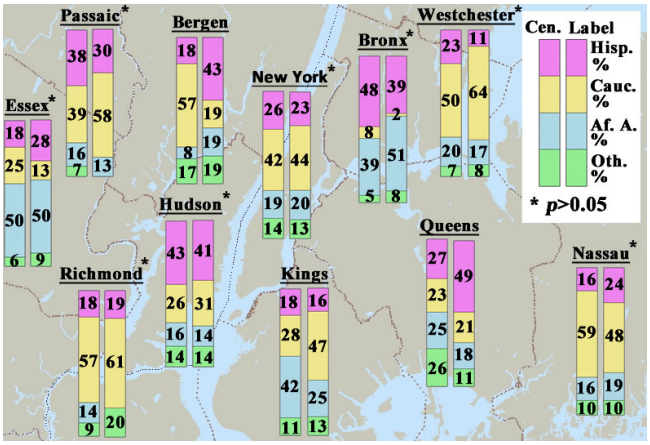
Figure 3: Daily ranges in miles. Top: boxes show the 25th, 50th, and 75th percentiles; whiskers the 2nd and 98th percentiles. Bottom: table with the percentiles represented in the boxplots. The maximum range (Max. Mo.–Fr.) is a user's longest distance and the median range (Med. Mo.–Fr.) a user's median distance, each taken on a single day for the entire Spring subset on a weekday [21]. 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 [20]. Previous results [20, 21] are shown in parentheses. 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. We define $\epsilon < 0.005$ miles.

4.3 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. In particular, we examine differences in daily ranges, home ranges, and temporal checkin characteristics.

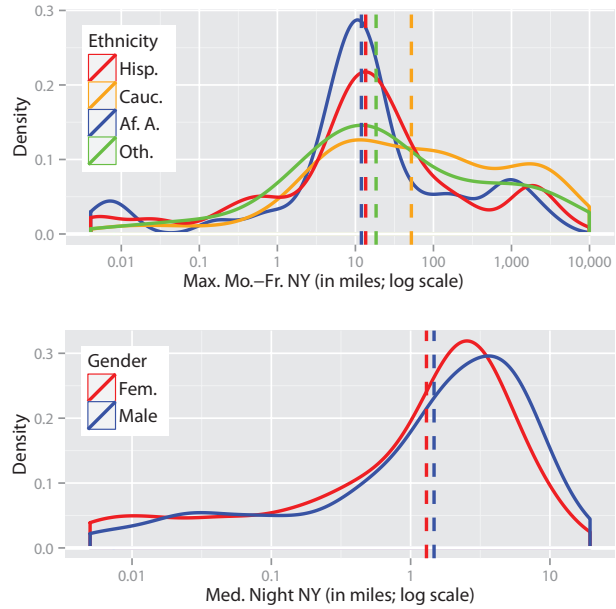
Daily Ranges.

Figure 5 shows some of our daily range results for ethnic groups and genders based on our sets of labeled users for LA and NY. We obtained the same types of daily ranges as described earlier in Figure 3, however, this time for all days of the year. It is striking that Caucasians generally have a higher 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 and 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.



	Ethnicity Multi-Cat.		*Ethnicity Binary*		*Gender*
Gran.	LA	NY	LA	NY	NY
State	0/1 (0%)	0/1 (0%)	1/1 (100%)	0/1 (0%)	1/1 (100%)
County	1/2 (50%)	**8/11 (73%)**	2/2 (100%)	6/8 (75%)	4/4 (100%)
PUMA	12/16 (75%)	11/17 (65%)	2/2 (100%)	5/6 (83%)	1/1 (100%)
NTA	- -	9/16 (56%)	- -	7/7 (100%)	2/2 (100%)
ZIP	3/3 (100%)	8/14 (57%)	1/1 (100%)	3/3 (100%)	- -

Figure 4: Chi square goodness of fit test results for ethnicity and gender at various levels of Census-defined granularity. Top: detailed view of the multi-category ethnicity distributions for the NY county level. Left bars show the Census distributions (Cen.) and right bars the label distributions (Label). Bottom: complete results of the chi square tests. NTAs are specific to NY and not available for LA. Below the ZIP code and NTA levels we did not have enough data to perform chi square tests. We follow [42] 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 also use a Monte Carlo simulation with 2,000 replicates.



	Max. Mo.-Fr. NY				*Med. Night NY*	
%	Hisp.	Cauc.	Af. A.	Oth.	Fem.	Male
98	2,480.8	6,509.4	2,270.9	6,788.1	9.8	11.5
75	50.8	592.3	44	187	3.2	4.7
50	**13.5**	**52.1**	**11.9**	**18.4**	**1.8**	**1.9**
25	4.9	7	5.5	3.7	0.4	0.6
02	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ

Figure 5: Daily ranges in miles. Top: density plot of the maximum daily ranges by ethnicity. Middle: density plot of the median daily ranges at night by gender. Bottom: table with the percentiles of the daily ranges represented in the plots. We rounded extremely small daily ranges up to 0.005 miles. 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. We define $\epsilon < 0.005$ miles.

The differences in ranges by ethnicity can be most prominently observed in the comparisons of Caucasians to African Americans and to Hispanics. However, it should be noted that at night all ethnicities exhibit 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 male and female ranges occurs for the median daily ranges at night. As shown in Figure 5, females tend to travel smaller distances at night than males. There are many possible explanations for this phenomenon. One reason could be that women travel fewer times at night due to safety concerns [2] and, consequently, 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 mobility with respect to an individual'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 travels. Different from daily ranges we calculate the home ranges not on a daily basis, but instead consider all home ranges—whether they were the maximum travel distance for a day or not. Based on a user's home location, as specified in §3, we calculate the distance between the home and each checkin for the different ethnic groups and genders. Figure 6 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 (most notably LA (2,475 miles), San Francisco (2,563 mi), and Seattle (2,405 miles)). Males and females have very similar home ranges at the edges of the graph. However, females travel farther in the medium home ranges. This finding could be based on the fact that women generally take more often vacations [22] and travel longer distances to work when they are employed full-time [24]. It should be noted that the larger home ranges are not inconsistent with the previous observation of shorter ranges for females at night as that result does obviously 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.

Temporal Checkin Characteristics.

Beyond spatial differences we explore differences in temporal activity as well. Figure 7 shows histograms for checkins by hour of day. As might be expected, we observe periodic behaviors with low checkin levels between 4–6am and peak levels from 3–8pm. On weekends the lows occur at later times than on weekdays suggesting that users wake up later on weekends. We also see a dramatic increase in activity after 5pm on weekdays, which could correspond

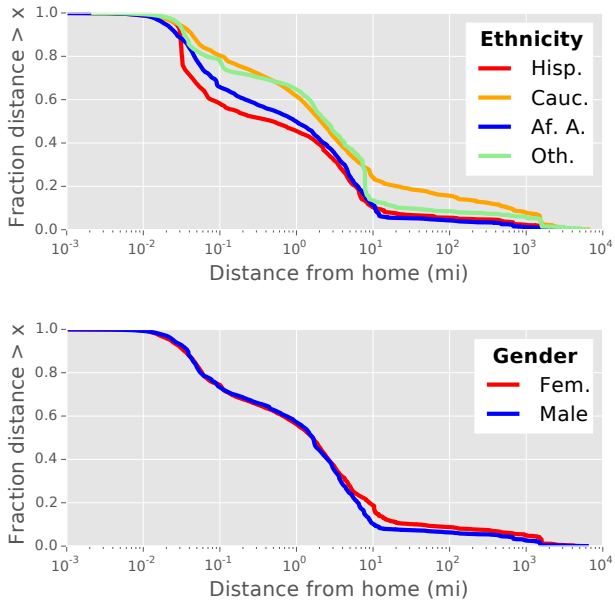


Figure 6: CCDFs of home ranges for NY. Top: CCDFs for different ethnic groups. Bottom: CCDFs for males and females.

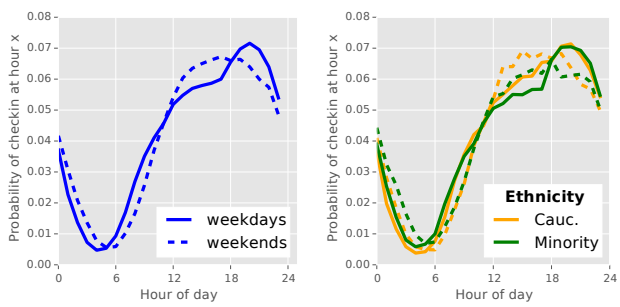


Figure 7: Histograms of checkin times for NY. Left: Comparison of weekends and weekdays for all user groups. Right: Comparison of Caucasian and minority user groups for weekends and weekdays. Dashed lines correspond to weekends, solid lines to weekdays.

to the time at which many users get off of work. When broken up into Caucasians and minorities, we see fairly similar curves, except with a more pronounced weekday after-work increase for minorities. It could be the case that Caucasians work more often in flexible environments. We observe no substantial differences between genders or NY and LA.

4.4 Ethnic Segregation

Location data are the basis for measuring residential segregation, that is, the degree to which two or more groups live separately from one another in different parts of the urban environment [33]. Trends in residential segregation characterize a group's proximity to community resources (e.g., health clinics) and its exposure to environmental and social hazards (e.g., poor water quality and crimes) [41]. In addition to *residential* segregation we also introduce and evaluate *mobility* segregation, which we understand as the degree to which two or more groups *move* to and from different parts of an area. Mobility segregation allows for a dynamic view of segregation, for example, in order to determine a group's ease of access to community resources away from home.

Methodology.

Various intersecting dimensions of segregation can be distinguished [33]. We explore two standard measures, each for a different dimension: the interaction index measures the dimension of exposure (the extent to which minority group members are exposed to majority group members in an area [33]) and the entropy index measures the dimension of evenness (the extent to which minority group members are over- or underrepresented in an area [33]). The interaction index, B, can be understood as the probability of a minority group member interacting with a majority group member and is defined [48] by

$$B_{kl} = \sum (\frac{n_{ik}}{N_k})(\frac{n_{il}}{n_i}), \tag{1}$$

where n_{ik} is the population of ethnic minority group k in area i (e.g., in a ZIP code area), N_k is the number of persons in group k in the total population of all areas, n_{il} is the population of ethnic majority group l in area i, and n_i is the area population.

The entropy index was used in social network research before [8] and has the advantage over other indices that it can be used to measure segregation for more than two groups. We define the entropy index [48], H, as

$$H = \frac{H^* - \bar{H}}{H^*}, \tag{2}$$

where H^* is the population-wide entropy defined by

$$H^* = -\sum_{k=1}^{K} P_k ln(P_k), \tag{3}$$

and $\bar{H}$ is the weighted average of the individual areas' entropies defined by

$$\bar{H} = -\sum_{i=1}^{I} \frac{n_i}{N} \sum_{k=1}^{K} P_{ik} ln(P_{ik}), \tag{4}$$

where K is the number of different ethnic groups, P_k is the proportion of ethnicity k in the total population, I is the number of different areas, n_i is the population in an area, N is the sum of the population from all areas, and P_{ik} is the proportion of the population of ethnicity k in area i (while it is defined that $P_{ik} ln(P_{ik}) = 0$ for $P_{ik} = 0$).

For both interaction and entropy indices we make use of our sets of labeled users for LA and NY, however, exclude all areas for which the label distribution deviated significantly from the Census distribution as indicated by $p \leq 0.05$. Thus, for example, as shown in Figure 4, on the county level we do not include Queens, Kings, and Bergen. These exclusions are necessary as otherwise the accuracy of our results decreases substantially. Recall that we define a user's home as the ZIP code where he or she had the most checkins (§3) and that we adjust label and Census distributions (§4.2).

Residential Segregation.

Tables 2 and 3 show our results for the interaction and entropy indices, respectively. For the most part the interaction between Caucasian and minority group members can be considered fairly high [18]. All three minorities in LA and NY have similar probabilities of interacting with Caucasians. The measurement errors of 5% (Hisp./Cauc. and Oth./Cauc.) and 6% (Af. A./Cauc.) between our labeled data and the Census suggest that our results are overall reliable. The inaccurate results for LA on the ZIP code level appear

	Hisp./Cauc.		*Af. A./Cauc.*		*Oth./Cauc.*	
Gran.	LA	NY	LA	NY	LA	NY
County	0.29 (-2%)	0.34 (+2%)	0.27 (+1%)	0.3 (-2%)	0.3 (-3%)	0.4 (0%)
PUMA	0.32 (-6%)	**0.39 (+3%)**	0.43 (+4%)	0.42 (+7%)	0.31 (-10%)	0.49 (+5%)
NTA	- -	0.54 (+6%)	- -	0.43 (+3%)	- -	0.55 (+7%)
ZIP	0.36 (-19%)	0.56 (0%)	0.33 (-23%)	0.55 (+1%)	0.58 (-1%)	0.5 (-7%)
∅ % Diff.	**5%**		**6%**		**5%**	

Table 2: Interaction index (B) for different granularities based on labeled Instagram data. Differences to the interaction index calculated from Census data are shown in percentage points in parenthesis. For example, the probability of a Hispanic person to interact with a Caucasian person on the PUMA granularity level for NY is 39%. However, as shown in parenthesis, this result is an overestimation by three percentage points over the Census distribution probability of 36%. The last row of the table shows the mean difference between our labels and the Census for the three different ethnicities in absolute percentage points for both LA and NY together. Note that NTAs are not available for LA and that we also did not analyze the state level as the label and Census distributions differ significantly (Figure 4).

	Entropy				
Metro	County	PUMA	NTA	ZIP	∅ % Diff.
LA	0.01 (-2%)	0.15 (+8%)	- -	0.15 (+9%)	**3%**
NY	0.08 (0%)	0.14 (+1%)	0.08 (0%)	0.09 (+4%)	

Table 3: Entropy index (H) for different granularities based on labeled Instagram data. Differences to the entropy index calculated from Census data are shown in percentage points in parenthesis. As explained in Table 2, the last column shows the measurement error. As further explained in Table 2, we did not consider NTA (LA) and state granularities (LA and NY).

to have been caused by the smaller number of data points. While the level of interaction seems to increase when areas become more fine-grained, this phenomenon seems to be caused by the different area coverage for the various granularities. For example, it is not present when considering all NY city areas, where the Census distributions for the interaction of African Americans and Caucasians are: 0.41 (County), 0.25 (PUMA), 0.2 (NTA), and 0.22 (ZIP).

With entropy index scores ranging from 0.01 to 0.15, as shown in Table 3, we find another indicator for low segregation [18]. However, it should be noted that this low level of segregation is a characteristic of the particular areas we investigated. For example, for all NY city areas at the NTA level we calculated an entropy of 0.31 indicating higher segregation. However, with mean differences of 5% (Hisp./Cauc.) and 6% (Af. A./Cauc. and Hisp./Oth.) between the results for our labeled data and the Census-based calculation our findings are generally reliable. As in the case of interaction, we believe that any existing inaccuracies could be due to small numbers of data points.

Mobility Segregation.

We evaluate mobility segregation based on the same measures as residential segregation—interaction and entropy indices. However, instead of using home locations we leverage checkin data. More specifically, for each user we calculate the percentage that he or she spent at a certain area and sum the resulting values for all users of a certain ethnicity. This method aims to avoid overcounting of active users. Our results are shown in Table 4 and indicate that segregation levels in terms of where people go are similar to levels of where people live. Indeed, it would have been surprising to see higher segregation levels as members of minority groups may work in predominantly Caucasian areas. Furthermore, it would also have been a surprise to see lower levels of segregation as residential segregation is already relatively low.

	Interaction			*Entropy*
Metro	Hisp./Cauc.	Af. A./Cauc.	Oth./Cauc.	All Eth.
LA	0.55 (+1%)	0.57 (0%)	0.58 (-1%)	0.06 (+1%)
NY	0.54 (-2%)	0.53 (-1%)	0.53 (-5%)	0.06 (+2%)
∅ % Diff.	**1%**	**1%**	**3%**	**1%**

Table 4: Mobility interaction and entropy indices for ZIP code granularity based on labeled Instagram checkin data. Differences to the residential interaction and entropy indices calculated from Census data are shown in percentage points in parenthesis. The last row of the table shows the mean difference between our labels and the Census in absolute percentage points for both LA and NY together.

5. INFERENCES FROM MOBILITY DATA

We now show how location data by itself allows to infer ethnicity and gender of individual Internet users. We introduce a simple frequentist approach (§5.1), describe considerations informing our methodology (§5.2), and present the results of its application (§5.3).

5.1 A Simple Inference Algorithm

Our approach yields two advantages: (1) it provides a formulation of the problem that is intuitive and (2) it remains generic so as to be easily applicable to any sparse location dataset. We use the following assumptions: each user, i, belongs to one of two classes, C_1 or C_2. Class C_1 (respectively C_2) is associated with a probability distribution μ_1 (respectively μ_2) over a discrete set of locations, representing the fraction of time spent by users of that class in that location. Our main assumption is that a user i makes n checkins, denoted $X^{(i)} = (X_1^{(i)}, \ldots, X_n^{(i)})$ at locations that are drawn independently from this user's class probability distribution. The prior probability that a user is in class C_1 or C_2 is denoted π_1 and π_2, respectively.

Note that this model does not use notions of times of the day, geographies, or auxiliary information. It applies to most location datasets as it is agnostic to how they were generated, anonymized, or in which granularity they are available. Such model serves as a starting point to approximate human mobility [15]. However, in practice humans show periodicity [14] or even social bias [7] in their movements, and users in a class may not be identically distributed, which is why it is important to test our technique using real data (§5.3). Under our assumptions, the problem of classifying users in their respective class reduces to a simple hypothesis testing. If i is in class C_1 then for any location l, we have

$$\forall j,\ P(X_j^{(i)} = l | i \in C_1) = \mu^{(1)}(l), \tag{5}$$

so that

$$P(X^{(i)} = (l_1, \ldots, l_n) | i \in C_1) = \mu^{(1)}(l_1) \ldots \mu^{(1)}(l_n), \tag{6}$$

by independence, and applying Bayes' rule

$$P(i \in C_1 | X^{(i)} = (l_1, \ldots, l_n)) = \frac{1}{1 + \frac{\pi_2 \mu^{(2)}(l_1) \ldots \mu^{(2)}(l_n)}{\pi_1 \mu^{(1)}(l_1) \ldots \mu^{(1)}(l_n)}}. \tag{7}$$

The Neyman-Pearson lemma states under the assumptions above that the most powerful statistical test to determine which class a user belongs to from its checkins is the likelihood ratio test. A maximum likelihood rule classifies a user in class 1 iff

$$\pi_2 \mu^{(2)}(l_1) \ldots \mu^{(2)}(l_n) < \pi_1 \mu^{(1)}(l_1) \ldots \mu^{(1)}(l_n) \tag{8}$$

Task	*Best Algorithm*	*Parameters*	*Important Features*	*Baseline Accuracy*	*Accuracy*	*AUC*	*F1*
Ethnicity NY	Logistic Regression	L1, $C = 0.01$	Avg. ZIP ethnicities	0.52	**0.72**	**0.76**	**0.74**
Ethnicity LA	Logistic Regression	L1, $C = 1$	Avg. ZIP ethnicities	0.50	**0.63**	**0.66**	**0.64**
Gender NY	Logistic Regression	L2, $C = 0.1$	Men's Store	0.53	**0.58**	**0.59**	**0.55**

Table 5: Results for the binary classifications of ethnicity and gender in NY and LA. The algorithms ran on all available features, such as counts of visits to different neighborhoods, the ethnicity of the most visited block, and the categories of nearby Foursquare venues. The baseline was obtained by predicting the class of a user based on the label distribution.

or, equivalently, if we have

$$\sum_{k=1}^{n} \ln \frac{\mu^{(1)}(l_k)}{\mu^{(2)}(l_k)} > \ln \frac{\pi_2}{\pi_1} \ . \tag{9}$$

We expect that our predictions are more accurate on users with more checkins. One can show under these assumptions that this classifier's error probability for a user decreases *exponentially* as the number of checkins n grows, that is,

$$P(\text{error}|n \text{ checkins}) \approx_{n\to\infty} 2^{-n\mathcal{C}(\mu_1,\mu_2)}, \tag{10}$$

where μ_1 and μ_2 are the probability distributions associated with C_1 and C_2, and $\mathcal{C}$ denotes the *Chernoff information*, defined as $\mathcal{C}(\mu_1, \mu_2) = -\min_{0\le\lambda\le 1} \ln \sum_l \mu_1(l)^{1-\lambda}\mu_2(l)^{\lambda}$.

Based on this analysis, a simple algorithm to infer ethnicity or gender can first estimate μ_1, μ_2 and π_1, π_2 using the training data and then classify according to this likelihood rule.

5.2 Methodology

Our purpose is to explore generally what might be inferred about users from their location data only. This affected our methodology in a few key ways. First, we utilized well-understood, commonly-applied techniques that could easily be employed by anyone with access to mobility data. We also used publicly available data-sources. Second, to make our results applicable to other sources of location data beyond Instagram, we did not use features specific to Instagram, such as the social network graph or user-generated descriptions. Thus, our work should be viewed as a lower-bound on the accuracy of what can be inferred using location data. Adversaries with access to more detailed auxiliary information, more data about each user (such as a contact list or recent purchases), or more advanced machine learning techniques might achieve better results.

We considered two questions: (1) Can minorities be distinguished from Caucasians? (2) Can women be distinguished from men? We represented users as feature vectors, using three classes of features: **geographic** features, such as counts or percentages of visits to locations; **semantic** features derived from Foursquare, such as the popularity of visited venues or counts of visits to venues with certain categories like "Restaurant" or "Park" (the collection of which we explained in §3); and **Census** derived features, such as the average ethnic makeup of all visited locations or the ethnic makeup of a user's most-visited location.

We performed all our experiments using the scikit-learn library [38] and tested the algorithms logistic regression, decision trees, naive Bayes, and support vector machines (SVMs). As a baseline, we predicted ethnicity or gender based on the class distribution, giving us baseline accuracies of 52% for ethnicity in NY, 50% for ethnicity in LA, and 53% for gender in NY.

Auxiliary Data.

Auxiliary information about a location derived from Foursquare or the Census may not always be available, e.g., in countries without publicly available census data or when locations are anonymized. Furthermore, a labeled training set of user data may not always be available either. To understand the performance of an algorithm that does not have access to any data beyond counts of visits to locations, we applied our **Bayesian** algorithm to our data. To test if labeled data was necessary to guess ethnicity, we developed a simple decision rule that used no labels. Based on Census data we calculated the average percentage of Caucasians living in all locations that a user visited. If this percentage was over the metropolitan area's average, we predicted that the user was Caucasian. If it was below, we predicted that the user was of a minority ethnicity. We called this the **Unsupervised Threshold** algorithm. We compared this algorithm to an algorithm with access to labeled data, which learned an optimal threshold rather than using one derived from publicly available Census data and which we dubbed the **Supervised Threshold** algorithm. Finally, we compared these algorithms against our best performing algorithm, run with all features at the lowest granularity. We call this the **Full** algorithm.

Data Granularity.

The granularity of location data can vary greatly depending on how it is created. Previous research has investigated the impact of location granularity on anonymity [9, 49]. To investigate the impact of granularity on inferences, we represented our location data at several different granularities defined by the Census ranging from block groups to states. The ethnic makeup of a large granularity area, such as a county, will typically be more similar to the overall metropolitan area's ethnic makeup than a small granularity area like a city block. Thus, increasing the granularity should make inferences more difficult.

Data Quantity.

Finally, with four different analyses, we studied the impact of data quantity on prediction accuracy. First, to explore the impact of user activity on inference accuracy, we grouped users according to their number of geolocated Instagram photos. Next, we investigated the impact of location diversity by grouping users according to the number of distinct ZIP codes they visited. Both of these are impacted by choices made by users—users who post more might be inherently easier to identify or predict. We thus did two more analyses where we sampled locations from a user's full set of checkins. In the first, we ran the Supervised Threshold algorithm on a user's k most visited locations. In the second, we ran the Supervised Threshold algorithm on n randomly sampled checkins.

5.3 Results

The results of our best-performing algorithms are displayed in Table 5, and a detailed comparison of accuracy as a function of granularity can be seen in Figure 8. Our results suggest that geotag data can be used to infer an individual's ethnicity and gender. The accuracy for predicting ethnicity falls squarely within what has been reported for other types of datasets. On the lower bound, in their work of predicting individual Twitter users as African American or not based on linguistic features of Tweets [39] report as best performance an F-1 score of 0.66. On the upper bound, for predicting whether the ethnic origin of a phone user is inside or outside the United States based on a rich feature set containing Internet

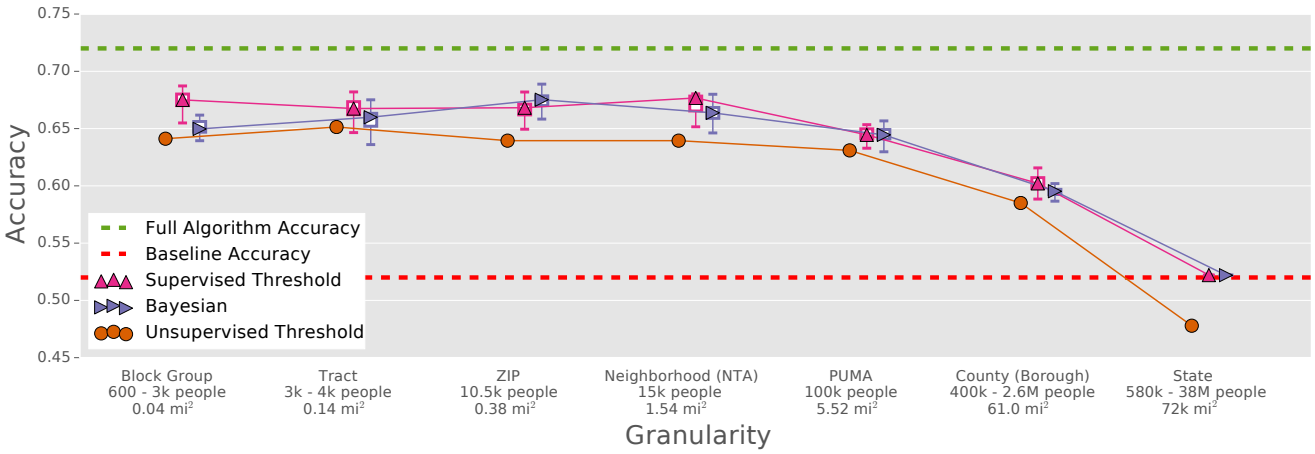


Figure 8: Accuracy of ethnicity prediction versus granularity for our NY population using several different inference techniques. Accuracy increases slightly at the ZIP code and neighborhood granularities and then decreases. Interestingly, the Bayesian algorithm, which uses only counts of visits to locations, performs comparably to the Supervised Threshold algorithm, which uses data on the ethnicity of visited locations.

usage, call, text message, and location features [1] achieved an F-measure of 0.81 and for gender an F-measure of 0.61. For gender [50] achieved an F-measure of 0.81 for social network users in Beijing and 0.82 for Shanghai based on spatial, temporal, and location context knowledge. Given that our dataset contains far fewer features our results demonstrate that geotags are surprisingly powerful in predicting gender and ethnicity.

Auxiliary Data.

It can be observed in Figure 8 that the Supervised Threshold algorithm performs much better than the Unsupervised Threshold algorithm suggesting that labeled data improves the algorithmic accuracy across the board by roughly 5%. Interestingly, the Bayesian algorithm performs comparably to the Supervised Threshold algorithm. Thus, an algorithm with no semantic information about visited locations performs just as well as one that knows the ethnic makeup of all visited locations. This suggests that an adversary with enough location data labeled with demographic data could obtain reasonable levels of accuracy with no knowledge of what locations were visited. Even if locations are "anonymized," that is, GPS coordinates or venue names were obscured, they can still be used to infer demographic information about the user.

Data Granularity.

The Full algorithm (that is, our best performing algorithm, with access to all features at all levels of granularity) achieves the best performance; no algorithm with access to restricted, coarser-grained features is as accurate.

The performance of all algorithms decreases at the most coarse granularities. This is most likely because the ethnicity distributions of larger regions are closer to the overall distribution of the metropolitan area and provide less information. Several algorithms improve in performance at medium granularities, such as ZIP and neighborhood. This is most likely caused by the sparsity of our dataset at the most detailed granularity as many blocks are only visited by a few users.

Data Quantity.

It appears that the accuracy of ethnicity prediction improves with the total number of checkins a user has made as shown in Figure 9. The distinct number of ZIP checkins of a user provides a separate measure of user activity as a user could have a large fraction of checkins in few ZIP codes. We can observe a substantial boost in accuracy after a user checked in at 12 distinct ZIP codes.

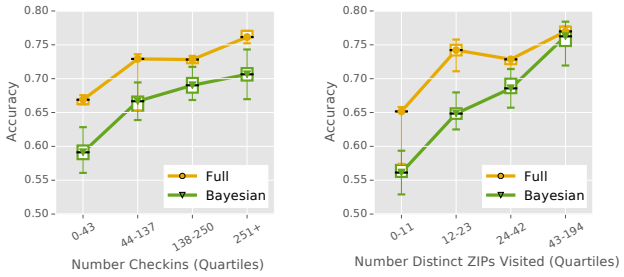


Figure 9: Checkin user activity. Left: accuracy as a function of total number of checkins at ZIP code locations. Right: accuracy as a function of number of checkins at distinct ZIP code locations.

We also found that when a user is only observed in a limited set of locations, the inference accuracy increases fast with a relatively small increase in the number of locations. Moreover, it is not even required to focus on the most significant locations of a user to get good inference accuracy. Observations of a user in a few random locations at the tract or neighborhood level might be enough for predicting ethnicity, and those locations may be even selected randomly and must not be necessarily related to the user's most significant places. These results, which are displayed in Figure 10, suggest that inference for the purpose of ethnicity identification is quite robust to data sparseness and obfuscation methods.

6. CONCLUSION

This study highlights the risks and opportunities of discriminative big data analysis by demonstrating that it is possible to infer Internet users' ethnicities and genders based on location data *alone*. It also shows that mobility patterns can be studied using publicly available data. Internet users may often be unaware that releasing such data could also disclose possibly sensitive personal information. Simply reducing granularity proved to be insufficient to prevent such privacy leakage as mobility remains discriminative. However, the trove of geotagged pictures available through individual online profiles also yields important insights for beneficial uses, for example, by city planners and social scientists.

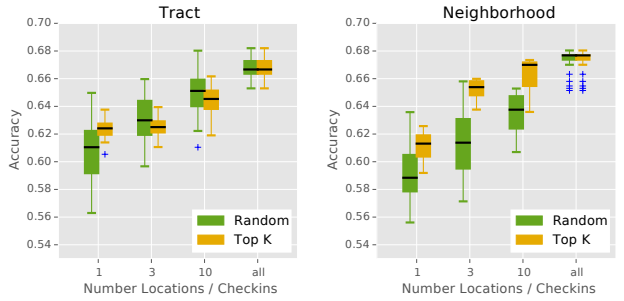


Figure 10: Accuracy of predicting a user's ethnicity from a small number of locations chosen either as most frequently visited locations or randomly. The algorithm used is the Supervised Threshold algorithm. Left: tract granularity. Right: neighborhood granularity.

As our dataset is similar, both demographically and mobility-wise, to other datasets as shown in §4, we believe that our results are generalizable and applicable to other unlabeled datasets. Although it could be claimed that our data is biased by the fact that the users in our study have willingly disclosed their gender and ethnicity by publicly using Instagram, we want to stress that it would be difficult and possibly unethical to create a labeled dataset of users who *do not* want to disclose their gender and ethnicity.

This work motivates multiple avenues of further research: First, it enables the extension of demographic mobility analysis to many researchers using shareable public datasets and reproducible results. Beyond ethnicity and gender, attributes such as age, occupation, and other lifestyle features may be extracted from users' pictures, and naturally there are many other mobility properties to account for beyond, for example, daily ranges. Second, better understanding the discriminative power of location data might inform the design of tools for raising user awareness about the information they reveal. This insight motivates revisiting mobility modeling and the inferences it renders possible to empower users to make at will their locations as clear as a photograph or as opaque as footprints in the mud.

7. ACKNOWLEDGMENTS

We would like to thank Danny Echikson and Stephanie Huang for their help labeling, Mathias Lécuyer for valuable discussion, and the anonymous reviewers for their useful suggestions.

8. REFERENCES

[1] Y. Altshuler, N. Aharony, M. Fire, Y. Elovici, and A. Pentland. Incremental learning with accuracy prediction of social and individual properties from mobile-phone data. In *SocialCom/PASSAT*, pages 969–974. IEEE, 2012.

[2] E. Badger. This is how women feel about walking alone at night in their own neighborhoods. http://www.washingtonpost.com/blogs/wonkblog/wp/2014-/05/28/this-is-how-women-feel-about-walking-alone-at-night-in-their-own-neighborhoods/, May 2014.

[3] R. Becker, R. Cáceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky. Human mobility characterization from cellular network data. *Communications of the ACM*, 56(1), Jan. 2013.

[4] J. Brea, J. Burroni, M. Minnoni, and C. Sarraute. Harnessing Mobile Phone Social Network Topology to Infer Users Demographic Attributes. In *SNAKDD'14: Proceedings of the 8th Workshop on Social Network Mining and Analysis*. ACM Request Permissions, Aug. 2014.

[5] J. Chang, I. Rosenn, L. Backstrom, and C. Marlow. epluribus: Ethnicity on social networks, 2010.

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

[7] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *KDD '11: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM Request Permissions, Aug. 2011.

[8] J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, UbiComp '10, pages 119–128, New York, NY, USA, 2010. ACM.

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

[10] Y.-A. de Montjoye, J. Quoidbach, F. Robic, and A. S. Pentland. Predicting personality using novel mobile phone-based metrics. In *Proceedings of the 6th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, SBP'13, pages 48–55, Berlin, Heidelberg, 2013. Springer-Verlag.

[11] Z. Deng and M. Ji. *Deriving Rules for Trip Purpose Identification from GPS Travel Survey Data and Land Use Data: A Machine Learning Approach*, chapter 72, pages 768–777. 2010.

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

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

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

[15] M. Grossglauser and D. Tse. Mobility increases the capacity of ad hoc wireless networks. *Networking, IEEE/ACM Transactions on*, 10(4):477–486, 2002.

[16] S. Guha, M. Jain, and V. N. Padmanabhan. Koi: a location-privacy platform for smartphone apps. In *NSDI'12: Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*. USENIX Association, Apr. 2012.

[17] Y. Hu, L. Manikonda, and S. Kambhampati. What we instagram: A first analysis of instagram photo content and user types, 2014.

[18] J. Iceland, D. Weinberg, and L. Hughes. The residential segregation of detailed Hispanic and Asian groups in the United States: 1980-2010. *Demographic Research*, 3:593–624, 2014.

[19] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky. Identifying important places in people's lives from cellular network data. *Pervasive Computing*, pages 133–151, 2011.

[20] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky. Ranges of human mobility in Los Angeles and New York. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on*, pages 88–93, 2011.

[21] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, J. Rowland, and A. Varshavsky. A tale of two cities. In *HotMobile '10: Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications*. ACM Request Permissions, Feb. 2010.

[22] Kelton. 4th annual springhill suites annual travel survey. http://news.marriott.com/springhill-suites-annual-travel-survey.html, April 2013.

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

[24] M.-P. Kwan. Gender, the home-work link, and space-time patterns of nonemployment activities. *Economic Geography*, 75(4):pp–370, 1999.

[25] N. Lathia, D. Quercia, and J. Crowcroft. The hidden image of the city: Sensing community well-being from urban mobility. In J. Kay, P. Lukowicz, H. Tokuda, P. Olivier, and A. Krüger, editors, *Pervasive*, volume 7319 of *Lecture Notes in Computer Science*, pages 91–98. Springer, 2012.

[26] K. Lewis, J. Kaufman, and N. Christakis. The taste for privacy: An analysis of college student privacy settings in an online social network. *J. Computer-Mediated Communication*, 14(1):79–100, 2008.

[27] L. Liao, D. Fox, and H. Kautz. Extracting places and activities from GPS traces using hierarchical conditional random fields. *Int. J. Rob. Res.*, 26(1):119–134, Jan. 2007.

[28] J. Lindamood, R. Heatherly, M. Kantarcioglu, and B. Thuraisingham. Inferring private information using social network data. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 1145–1146, New York, NY, USA, 2009. ACM.

[29] F. Liu, D. Janssens, G. Wets, and M. Cools. Annotating mobile phone location data with activity purposes using machine learning algorithms. *Expert Syst. Appl.*, 40(8):3299–3311, June 2013.

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

[31] M. Madden, A. Lenhart, S. Cortesi, U. Grasser, M. Duggan, A. Smith, and M. Beaton. Teens, social media, and privacy. *Pew Research Center*, 2013.

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

[33] D. S. Massey and N. A. Denton. The dimensions of residential segregation. *Social Forces*, 67(2):281–315, 1988.

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

[35] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist. Understanding the Demographics of Twitter Users. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11)*, Barcelona, Spain, July 2011.

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

[37] G. Paolacci, J. Chandler, and P. G. Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision Making*, 5(5):411–419, 2010.

[38] F. Pedregosa et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[39] M. Pennacchiotti and A.-M. Popescu. A machine learning approach to twitter user classification, 2011.

[40] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents*, SMUC '10, pages 37–44, New York, NY, USA, 2010. ACM.

[41] S. F. Reardon. *A Conceptual Framework for Measuring Segregation and its Association with Population Outcomes*, chapter 7, pages 169–192. John Wiley Sons, San Francisco, CA, USA, 2006.

[42] J. T. Roscoe and J. A. Byars. 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, Dec. 1971.

[43] C. Sarraute, P. Blanc, and J. Burroni. A study of age and gender seen through mobile phone usage patterns in Mexico. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 836–843, 2014.

[44] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[45] Statista. Social networking time per user in the united states in july 2012, by ethnicity (in hours and minutes). http://www.statista.com/statistics/248158/social-networking-time-per-us-user-by-ethnicity/, 2012.

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

[47] United States v. Jones. 2012. 132 S. Ct. 945, 955 (Sotomayor, J., concurring) (quoting People v. Weaver, 12 N.Y.3d 433, 441-42 (2009)).

[48] M. J. White. Segregation and diversity measures in population distribution. *Population Index*, 52(2):198–221, 1986.

[49] 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.

[50] 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.