Understanding the Demographics of Twitter Users

Alan Mislove[†] **Sune Lehmann**[*] **Yong-Yeol Ahn**[†] **Jukka-Pekka Onnela**[‡] **J. Niels Rosenquist**[‡]

[†]Northeastern University [*]Technical University of Denmark [‡]Harvard Medical School

Abstract

Every second, the thoughts and feelings of millions of people across the world are recorded in the form of 140-character tweets using Twitter. However, despite the enormous potential presented by this remarkable data source, we still do not have an understanding of the Twitter population itself: Who are the Twitter users? How representative of the overall population are they? In this paper, we take the first steps towards answering these questions by analyzing data on a set of Twitter users representing over 1% of the U.S. population. We develop techniques that allow us to compare the Twitter population to the U.S. population along three axes (geography, gender, and race/ethnicity), and find that the Twitter population is a highly non-uniform sample of the population.

Introduction

Online social networks are now a popular way for users to connect, communicate, and share content; many serve as the de-facto Internet portal for millions of users. Because of the massive popularity of these sites, data about the users and their communication offers unprecedented opportunities to examine how human society functions at scale. However, concerns over user privacy often force service providers to keep such data private. Twitter represents an exception: Over 91% of Twitter users choose to make their profile and communication history publicly visible, allowing researchers access to the vast majority of the site. Twitter, therefore, presents a unique opportunity to examine the public communication of a large fraction of the population.

In fact, researchers have recently begun to use the content of Twitter messages to measure and predict real-world phenomena, including movie box office returns (Asur and Huberman 2010), elections (O'Connor et al. 2010), and the stock market (Bollen, Mao, and Zeng 2010). While these studies show remarkable promise, one heretofore unanswered question is: *Are Twitter users a representative sample of society?* If not, which demographics are over- or underrepresented in the Twitter population? Because existing studies generally treat Twitter as a "black box," shedding light on the characteristics of the Twitter population is likely to lead to improvements in existing prediction and measurement methods. Moreover, understanding the characteristics of the Twitter population is crucial to move towards more advanced observations and predictions, since such an understanding will help us determine what predictions can be made and what other data is necessary to correct for any biases.

In this paper, we take a first look at the demographics of the Twitter users, aiming to answer these questions. To do so, we use a data set of over 1,755,925,520 Twitter messages sent by 54,981,152 users between March 2006 and August 2009 (Cha et al. 2010). We focus on users whose identified location is within the United States, because the plurality of users at the time of the data collection are in U.S., and because we have the detailed demographic data for U.S. population. Even with the location constraint, our dataset covers over three million users, representing more than 1% of the entire U.S. population.

Ideally, when comparing the Twitter population to society as a whole, we would like to compare properties including socio-economic status, education level, and type of employment. However, we are restricted to only using the data that is (optionally) self-reported and made visible by the Twitter users, including their name, location, and the text of their tweets. We develop techniques to examine the properties of the Twitter population along three separate but interrelated axes, based on the feasibility of comparison. First, we compare the geographic distribution of users to the population as a whole using U.S. Census data. We demonstrate that Twitter users are more likely to live within populous counties than would be expected from the Census data, and that sparsely populated regions of the U.S. are significantly underrepresented. Second, we infer the gender of Twitter users and demonstrate that a significant male bias exists, although the bias is becoming less pronounced over time. Third, we examine the race/ethnicity of Twitter users and demonstrate that the distribution of race/ethnicity is highly geographically-dependent.

Geographic distribution

Detection location using self-reported data

To determine geographic information about users, we use the self-reported *location* field in the user profile. The location is an optional self-reported string; we found that 75.3% of the publicly visible users listed a location. In order to turn the user-provided string into a mappable location, we use the Google Maps API. Beginning with the most popular location strings (i.e, the strings provided by the most users), we query Google Maps with each location string. If Google Maps is able to interpret a string as a location, we receive a latitude and longitude as a response. We restrict our scope to users in the U.S. by only considering response latitudes and longitudes that are within the U.S.. In total, we find mappings to a U.S. longitude and latitude for 246,015 unique strings, covering 3,279,425 users (representing 8.8% of the users who list a location).

To compare our Twitter data to the 2000 U.S. Census, it is necessary to aggregate the users into U.S. counties. Using data from the U.S. National Atlas and the U.S. Geological Survey, we map each of the 246,015 latitudes and longitudes into their respective U.S. county. Unless otherwise stated, our analysis for the remainder of this paper is at the U.S. county level.

Limitations We now briefly discuss potential limitations of our location inference methodology. First, it is worth noting that Google Maps will also interpret locations that are at a granularity coarser than a U.S. county (e.g., "Texas"). We manually removed these, including the mappings of all 50 states, as well as "United States" and "Earth." Second, users may lie about their location, or may list an out-of-date location. Third, since the location is per-user (rather than per-tweet), a user who moves from one city to another (and updates his location) will have all of his tweets considered as being from the latter location.

Geographic distribution of Twitter users

We begin by examining the geographic distribution of Twitter users, and comparing it to the entire U.S. population. Overall, the 3,279,425 Twitter users who we are able to geolocate represent 1.15% of the entire population (at the time of the 2000 Census). However, if we examine the distribution of Twitter users per county, we observe a highly non-uniform distribution.

Figure 1 presents this analysis, with the county population along the x axis and the fraction of this population we observe in Twitter along the y axis. We see that, as the population of the county increases, the *Twitter representation rate* (simply the number of Twitter users in that county divided by the number of people in that county in the 2000 U.S. Census) increases as well. For example, consider the median per-county Twitter representation rate of 0.324%. We observe that 93.5% of the counties with over 100,000 residents have a higher Twitter representation rate than the median, compared to only 40.8% of the counties with fewer than 100,000 residents (were Twitter users a truly random population sample, we would expect these percentages to both be 50%). Thus, the Twitter users significantly overrepresent populous counties, a fact underscored by the difference between the median (0.324%) per-county Twitter representation rates and the overall population sample of 1.15%.

The overrepresentation of populous counties in and of itself may not come as a surprise, due to the patterns of social media adoption across different regions. However, the

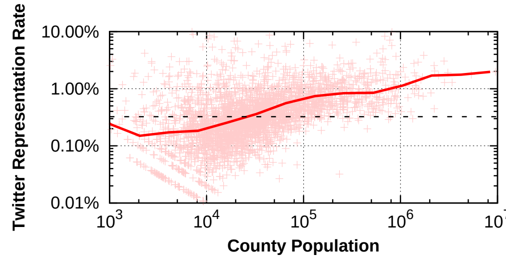


Figure 1: Scatterplot of US county population versus Twitter representation rate in that county. The dark line represents the aggregated median, and the dashed black line represents the overall median (0.324%). There is a clear overrepresentation of more populous counties.

magnitude of the difference is striking: We observe an order of magnitude difference in median per-county Twitter representation rate between counties with 1,000 people and counties with 1,000,000 people. This indicates a bias in the Twitter population (relative to the U.S. population) and suggests that entire regions of the U.S. may be significantly underrepresented.

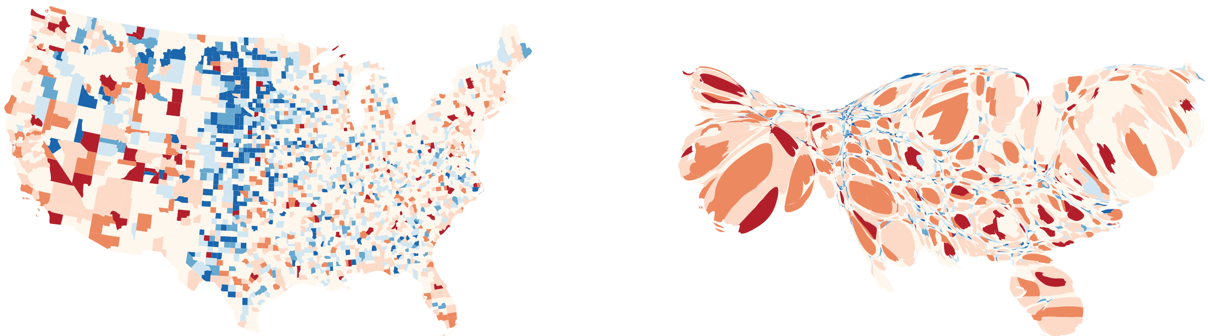
Distribution across counties We now examine *which* regions of the U.S. contain these over- and underrepresented counties. To do so, we plot a map of the U.S. based on the Twitter representation rate, relative to the median rate of 0.324%. Figure 2 presents this data, using both a normal representation and an area cartogram representation (Gastner and Newman 2004). In this figure, the counties are colored according to the level of over- or underrepresentation, with blue colors representing underrepresentation and red colors representing overrepresentation, relative to the median rate of 0.324%. Thus, the same number of counties will be colored red as blue.

These two maps lead to a number of interesting conclusions: First, as evident in the normal representation, much of the mid-west is significantly underrepresented in the Twitter user base in this time period. Second, as evident in the significantly red hue of the area cartogram, more populous counties are consistently oversampled. However, the level of oversampling does not appear to be dependent upon geography: Both east coast and west coast cities are clearly visible (e.g., San Francisco and Boston), as well as mid-west and southern cities (e.g, Dallas, Chicago, and Atlanta).

Gender

Detecting gender using first names

As we have very limited information available on each user, we rely on using the self-reported name available in each user's profile in order to detect gender. To do so, we first obtain the most popular 1,000 male and female names for babies born in the U.S. for each year 1900–2009, as reported by the U.S. Social Security Administration (Social Security Administration 2010). We then aggregate the names together, calculating the total frequency of each of the resulting 3,034 male and 3,643 female names. As certain names occurred in both lists, we remove the 241 names that were



(a) Normal representation

(b) Area cartogram representation

Figure 2: Per-county over- and underrepresentation of U.S. population in Twitter, relative to the median per-county representation rate of 0.324%, presented in both (a) a normal layout and (b) an area cartogram based on the 2000 Census population. Blue colors indicate underrepresentation, while red colors represent overrepresentation. The intensity of the color corresponds to the log of the over- or underrepresentation rate. Clear trends are visible, such as the underrepresentation of mid-west and overrepresentation of populous counties.

less than 95% predictive (e.g., the name Avery was observed to correspond to male babies only 56.8% of the time; it was therefore removed). The result is a list of 5,836 names that we use to infer gender.

Limitations Clearly, this approach to detecting gender is subject to a number of potential limitations. First, users may misrepresent their name, leading to an incorrect gender inference. Second, there may be differences in choosing to reveal one's name between genders, leading us to believe that fewer users of one gender are present. Third, the name lists above may cover different fractions of the male and female populations.

Gender of Twitter users

We first determine the number of the 3,279,425 U.S.-based users who we could infer a gender for, based on their name and the list previously described. We do so by comparing the first word of their self-reported name to the gender list. We observe that there exists a match for 64.2% of the users. Moreover, we find a strong bias towards male users: Fully 71.8% of the the users who we find a name match for had a male name.

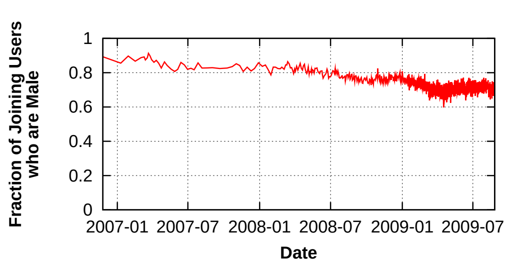


Figure 3: Gender of joining users over time, binned into groups of 10,000 joining users (note that the join rate increases substantially). The bias towards male users is observed to be decreasing over time.

To further explore this trend, we examine the historic gender bias. To do so, we use the join date of each user (available in the user's profile). Figure 3 plots the average fraction of joining users who are male over time. From this plot, it is clear that while the male gender bias was significantly stronger among the early Twitter adopters, the bias is becoming reduced over time.

Race/ethnicity

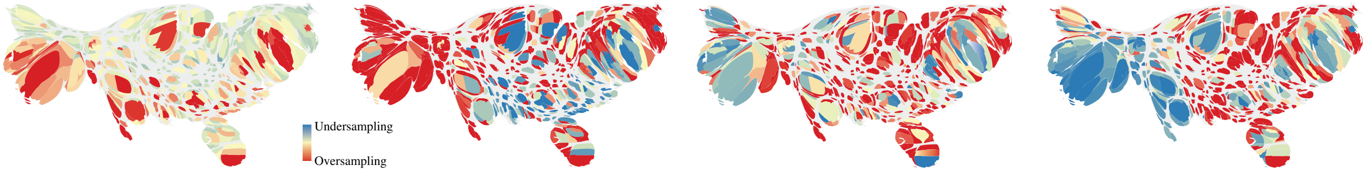
Detecting race/ethnicity using last names

Again, since we have very limited information available on each Twitter user, we resort to inferring race/ethnicity using self-reported last name. We examine the last name of users, and correlate the last name with data from the U.S. 2000 Census (U.S. Census 2000). In more detail, for each last name with over 100 individuals in the U.S. during the 2000 Census, the Census releases the distribution of race/ethnicity for that last name. For example, the last name "Myers" was observed to correspond to Caucasians 86% of the time, African-Americans 9.7%, Asians 0.4%, and Hispanics 1.4%.

Race/ethnicity distribution of Twitter users

We first determined the number of U.S.-based users for whom we could infer the race/ethnicity by comparing the last word of their self-reported name to the U.S. Census last name list. We observed that we found a match for 71.8% of the users. We the determined the distribution of race/ethnicity in each county by taking the race/ethnicity distribution in the Census list, weighted by the frequency of each name occurring in Twitter users in that county.[1] Due to the large amount of ambiguity in the last name-to-race/ethnicity list (in particular, the last name list is more than 95% predictive for only 18.5% of the users), we are unable to directly compare the Twitter race/ethnicity distribu-

[1]This is effectively the `census.model` approach discussed in prior work (Chang et al. 2010).



(a) Caucasian (non-hispanic) (b) African-American (c) Asian or Pacific Islander (d) Hispanic

Figure 4: Per-county area cartograms of Twitter over- and undersampling rates of Caucasian, African-American, Asian, and Hispanic users, relative to the 2000 U.S. Census. Only counties with more than 500 Twitter users with inferred race/ethnicity are shown. Blue regions correspond to undersampling; red regions to oversampling.

tion directly to race/ethnicity distribution in the U.S. Census. However, we are able to make *relative* comparisons between Twitter users in different geographic regions, allowing us to explore geographic trends in the race/ethnicity distribution. Thus, we examine the per-county race/ethnicity distribution of Twitter users.

In order to account for the uneven distribution of race/ethnicity across the U.S., we examine the per-county race/ethnicity distribution relative to the distribution from the overall U.S. Census. For example, if we observed that 25% of Twitter users in a county were predicted to be Hispanic, and the 2000 U.S. counted 23% of people in that county as being Hispanic, we would consider Twitter to be *oversampling* the Hispanic users in that county. Figure 4 plots the per-county race/ethnicity distribution, relative to the 2000 U.S. Census, per all counties in which we observed more than 500 Twitter users with identifiable last names. A number of geographic trends are visible, such as the undersampling of Hispanic users in the southwest; the undersamping of African-American users in the south and midwest; and the oversampling of Caucasian users in many major cities.

Related work

A few other studies have examined the demographics of social network users. For example, recent studies have examined the ethnicity of Facebook users (Chang et al. 2010), general demographics of Facebook users (Corbett 2010), and differences in online behavior on Facebook and MySpace by gender (Strayhorn 2009). However, studies of general social networking sites are able to leverage the broad nature of the profiles available; in contrast, on Twitter, users self-report only a minimal set of information, making calculating demographics significantly more difficult.

Conclusion

Twitter has received significant research interest lately as a means for understanding, monitoring, and even predicting real-world phenomena. However, most existing work does not address the sampling bias, simply applying machine learning and data mining algorithms without an understanding of the Twitter user population. In this paper, we took a first look at the user population themselves, and examined the population along the axes of geography, gender, and race/ethnicity. Overall, we found that Twitter users significantly overrepresent the densely population regions of the U.S., are predominantly male, and represent a highly nonrandom sample of the overall race/ethnicity distribution.

Going forward, our study sets the foundation for future work upon Twitter data. Existing approaches could immediately use our analysis to improve predictions or measurements. By enabling post-hoc corrections, our work is a first step towards turning Twitter into a tool that can make inferences about the population as a whole. More nuanced analyses on the biases in the Twitter population will enhance the ability for Twitter to be used as a sophisticated inference tool.

Acknowledgements

We thank Fabricio Benevenuto and Meeyoung Cha for their assistance in gathering the Twitter data used in this study. We also thank Jim Bagrow for valuable discussions and his collection of geographic data from Google Maps. This research was supported in part by NSF grant IIS-0964465 and an Amazon Web Services in Education Grant.

References

Asur, S., and Huberman, B. 2010. Predicing the Future with Social Media. In *WI*.

Bollen, J.; Mao, H.; and Zeng, X.-J. 2010. Twitter mood predicts the stock market. In *ICWSM*.

Cha, M.; Haddadi, H.; Benevenuto, F.; and Gummadi, K. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. In *ICWSM*.

Chang, J.; Rosenn, I.; Backstrom, L.; and Marlow, C. 2010. ePluribus: Ethnicity on Social Networks. In *ICWSM*.

Corbett, P. 2010. Facebook Demographics and Statistics Report 2010. `http://www.istrategylabs.com/2010/01/facebook-demographics-and-statistics-report-2010-145-growth-in-1-year`.

Gastner, M. T., and Newman, M. E. J. 2004. Diffusion-based method for producing density-equalizing maps. *PNAS* 101.

O'Connor, B.; Balasubramanyan, R.; Routledge, B.; and Smith, N. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *ICWSM*.

Social Security Administration. 2010. Most Popular Baby Names. `http://www.ssa.gov/oact/babynames`.

Strayhorn, T. 2009. Sex Differences in Use of Facebook and MySpace among First-Year College Students. *Stud. Affairs* 10(2).

U.S. Census. 2000. Genealogy Data: Frequently Occurring Surnames from Census. `http://www.census.gov/genealogy/www/data/2000surnames`.