Evaluating the Representativeness in the Geographic Distribution of Twitter User Population

Junjun Yin
Social Science Research Institute
Pennsylvania State University
State College, PA, USA
jyin@psu.edu

Guangqing Chi
Dept. of Agricultural Economics,
Sociology, and Education
Pennsylvania State University
State College, PA, USA
gchi@psu.edu

Jennifer Van Hook
Dept. of Sociology and Criminology
Pennsylvania State University
State College, PA, USA
jxv21@psu.edu

ABSTRACT

Twitter data are becoming a Big Data stream and have drawn multidisciplinary interests to study population characteristics and social problems that cannot be measured well by traditional surveys. However, the use of Twitter data has been strongly resisted because of concerns about the representativeness of the population as we know little about the demographic characters of the users. It is critical to evaluate the extent to which Twitter users represent the population across different demographic groups. This study evaluates the representativeness and examines the geographic distributions of Twitter user population and its correspondence to the real population. By estimating Twitter user demographics for the contiguous U.S. in 2014, the preliminary results revealed both over- and under-representation of certain demographic groups against the real population at county-level. A representation index is used to help depict and assess the representativeness of Twitter samples geographically, which may help further studies to identify the determinants of biases.

KEYWORDS

Geo-tagged Tweets, Demographics, Bias, Representativeness, Geographic Distribution

1 Introduction

The skyrocketing growth of social media data provides significant opportunities for studying social problems and advance social sciences. Twitter offers one of the most rapidly growing and accessible Big Data streams and has drawn interests from multiple disciplines for understanding various population dynamics, such as urban studies, public health, and behavioral science. However, the use of Twitter data has been strongly resisted by social scientists because of concerns about the representativeness of the population as a whole and because we know little about the demographic characteristics of the users.

To address the selection bias, the first step in using Twitter data is to understand the demographics of Twitter users. This study incorporates a set of state-of-the-art methods to estimate the demographic characters of Twitter users. Further, we evaluate the extent to which Twitter users represent the population by different demographic groups and examine the representativeness regarding the geographic distributions of twitter user population and its correspondence to the real population. This analysis uses geo-tagged Twitter data collected for the contiguous U.S. in 2014 and examines the geographic distribution of the representativeness for different demographic groups at the county level.

2 Data and Methods

In this study, we used geo-tagged tweets collected over the contiguous U.S. from January 1st to December 30th, 2014 by using the Twitter Streaming API[1]. The whole data collection contains approximately 1.2 billion tweets from over 6.4 million Twitter accounts (counted based on unique Twitter user ids). After removing non-human Twitter user accounts and tourists [3], the total geo-tagged Twitter user population was reduced to approximately 835 million tweets produced by 3.78 million unique Twitter users.

The analysis of the geographic distribution of the Twitter user population and its correspondence to the real population is performed at county level. We assigned each geo-tagged tweet to a corresponding U.S. county (of 3,105 counties in the contiguous U.S.) [2]. The demographic characters (i.e., gender, age, and race/ethnicity) of each individual consists of 11 groups: females, males, age groups (20–24, 25–34, 35–44, 45–54, 55–64, and 65+), Hispanics, non-Hispanic Whites, and non-Hispanic Blacks, which is then aggregated to the corresponding county. The county-level population for year 2014 is obtained from the American Community Survey (ACS) estimates, including total population estimates, population counts for different age groups, and percentage of the three race/ethnicity groups of the population.

Though many methods were developed to estimate Twitter user demographic characters, few studies have used them collectively to achieve optimal estimates [1]. We combined a set of state-of-the-art techniques for this study as illustrated in Figure 1: Age is estimated using Microsoft Azure facial recognition[2]; Gender is estimated based on first name extracted from user profile matching to a first name database from Facebook profiles and the profile image using Microsoft Azure facial recognition; Race/ethnicity is estimated based on last name extracted from user profile matching to U.S. Census Bureau's surname database for

[1] https://dev.twitter.com/streaming/overview

[2] https://azure.microsoft.com/en-us/services/cognitive-services/face/

race/ethnicity. With these methods, we were able to identify 80% gender, 45% age, and 52% race/ethnicity of the Twitter users.

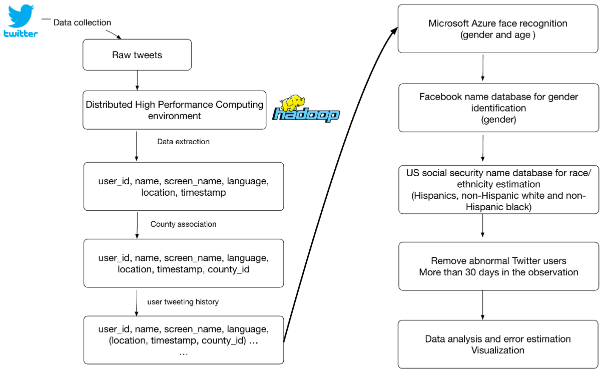
twitter
Data collection
Raw tweets
Distributed High Performance Computing environment
hadoop
Data extraction
user_id, name, screen_name, language, location, timestamp
County association
user_id, name, screen_name, language, location, timestamp, county_id
user tweeting history
user_id, name, screen_name, language, (location, timestamp, county_id) ... ...
Microsoft Azure face recognition (gender and age)
Facebook name database for gender identification (gender)
US social security name database for race/ ethnicity estimation (Hispanics, non-Hispanic white and non-Hispanic black)
Remove abnormal Twitter users More than 30 days in the observation
Data analysis and error estimation Visualization

Figure 1: The Flow Chart for Estimating Twitter User Demographics by Gender, Age, and Race/Ethnicity

To evaluate how representative Twitter users in each county are of the total population in each county, we created a representation (r) index, which is defined as:

$$r_j = (\frac{T_j}{T_{all}} / \frac{P_j}{P_{all}}) \qquad (1)$$

where, T_j denotes the number of Twitter users in county j, P_j denotes the population in county j, T_{all} denotes the total Twitter user population, and P_{all} denotes the total population. A value of $r_i = 1$ indicates that the percentage of Twitter users in county j is equal to the national average of Twitter users. $r_j > 1$ indicates an overrepresentation of Twitter users in county j, whereas $r_j < 1$ indicates an underrepresentation in county j.

3 Preliminary Results

In our preliminary study, the percentages of each Twitter user demographic group were compared to the corresponding ones from the ACS estimates. We measured the bias by the median percentage error and median absolute percentage error for each demographic group at the county level. As expected, Twitter users at ages 20–34 are over-represented while other age groups are under-represented. Non-Hispanic black Twitter users are also over-represented. The biases for gender and race/ethnicity are difficult to explain. This is partly due to the fact that the percentage error measures are sensitive to distributional features in the data (e.g., skewed distributions) and do not provide simple and intuitive descriptions of the biases. Therefore, we used the representation index to evaluate the geographic distribution of which Twitter users represent the population by different demographic groups. Figure 2(top) shows the representation index for the contiguous U.S. in 2014. Twitter users are overrepresented in metropolitan areas and underrepresented in rural areas. We can also apply Eq. (1) to each demographic group. Figure 2(bottom) shows the population index for Hispanics in 2014. Although Hispanics have a smaller percentage of the total population in the northeast, Hispanics in this region use Twitter more than those in the other regions.

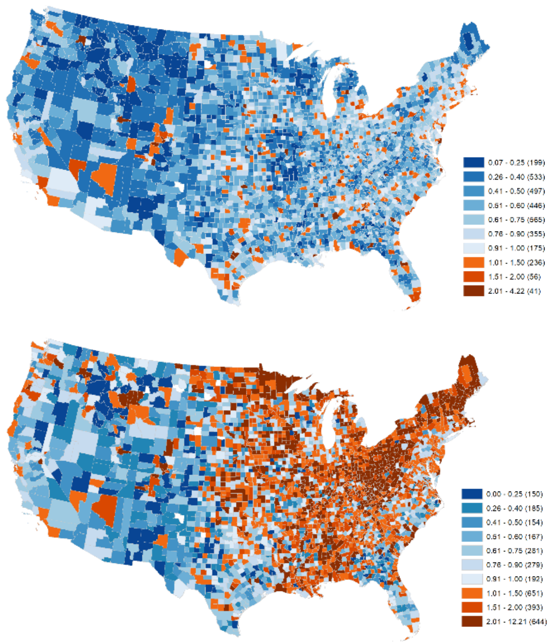

Figure 2. Representation Index for (top) All Twitter Users in 2014 and (bottom) Hispanics in 2014

4 Conclusions and Future Work

A major objective of this study is to provide knowledge to first understand the demographic characters of Twitter users and then evaluate the representativeness in the geographic distribution of twitter user population. We utilized a set of state-of-the-art methods to estimate Twitter user demographics for the contiguous U.S. in 2014. The analysis of the geographic distribution of the Twitter user population and its correspondence to the real population is performed at county level. The preliminary results have revealed both over- and under-representation of certain demographic groups regarding the real population at county-level. Further, the representation index helps to depict and assess the representativeness of Twitter samples geographically, which may help further studies to identify the determinants of biases.

One potential limitation is that only a small percentage of tweets are geo-tagged, so the geo-tagged tweets in this study may not be representative of all tweets. Future work will acquire non-geo-tagged tweets and infer their geolocations to reevaluate the representativeness. Importantly, such efforts will help adjust the weights in Twitter user samples for population related research.

REFERENCES

1. Feixiong Luo, Guofeng Cao, Kevin Mulligan, and Xiang Li. 2016. Explore spatiotemporal and demographic characteristics of human mobility via Twitter: A case study of Chicago. *Applied Geography* 70: 11–25.
2. Junjun Yin, Yizhao Gao, Zhenhong Du, and Shaowen Wang. 2016. Exploring Multi-Scale Spatiotemporal Twitter User Mobility Patterns with a Visual-Analytics Approach. *International Journal of Geo-Information* 5, 12: 187.
3. Junjun Yin, Aiman Soliman, Dandong Yin, and Shaowen Wang. 2017. Depicting urban boundaries from a mobility network of spatial interactions: a case study of Great Britain with geo-located Twitter data. *International Journal of Geographical Information Science* 31, 7: 1293–1313.