Visualizing Geo-Demographic Urban Data

Matheus Araujo
University of Minnesota
4-192 Keller H. 200 Union St SE,
Minneapolis, MN 55455
arauj021@umn.edu

Yelena Mejova
Qatar Computing Research Institute, HBKU
P.O. Box 34110, Doha, Qatar
yelenamejova@acm.org

Michaël Aupetit
Qatar Computing Research Institute, HBKU
P.O. Box 34110, Doha, Qatar
maupetit@hbku.edu.qa

Ingmar Weber
Qatar Computing Research Institute, HBKU
P.O. Box 34110, Doha, Qatar
iweber@hbku.edu.qa

CSCW'18 Companion, Nov. 2-7, 2018, Jersey City, NJ, USA.

ACM ISBN 978-1-4503-6018-0/18/11.
http://dx.doi.org/10.1145/3272973.3273001

Abstract
We present two interactive data visualizations of fine-grained demographic information for New York City, US, and Doha, Qatar, obtained using Facebook's Marketing API. The visualizations make innovative use of treemaps rather than typical checkboxes, to support informed filtering of data and visualization of both "where are people of type X" and "what type of people are in location Y". The two interactive visualizations aim to both show-case a front-end for census-type information and to demonstrate the richness of Facebook's advertising data.
Application URL: http://fb-doha.qcri.org/ and http://fb-nyc.qcri.org/.

Author Keywords
Visualization; demographics; online advertising; Facebook; New York City; Doha.

ACM Classification Keywords
Human-centered computing [Visualization]: Geographic visualization

Introduction
Fine-grained location data available in social media streams has recently been used to map activity [6] and detect neighborhoods [3] in a data-driven fashion. Extracting demographic information from social media is difficult, however,

as few users explicitly state their age, gender, and other personal attributes. So secondary tools including name dictionaries and image classifiers must be applied. However, an aggregated and anonymized, yet rich, source of demographic information is available through the advertising platforms of large social media sites, such as Facebook. Advertising data from Facebook's user base has been used to estimate stocks of migrants [7], track obesity [2] and visualize health interests around the world [1]. Here, we present a fine-grained visualization of multi-dimensional demographic data Facebook has to offer.

Data Collection, Processing and Validation

Facebook's marketing platform allows its advertisers to show advertisements to user cohorts matching specific demographic attributes[1]. These attributes can be explicitly self-declared by Facebook users, as in the case of gender or age, or they are inferred by the platform, as in the case of country of origin. Preliminary estimates of how many Facebook users match certain targeting criteria, the so-called "reach estimates", can be gathered using its Marketing API[2]. Here, we illustrate two geo-targeting methods that allow the visualization of a variety of demographic information available through this source, including gender, age group, education level, nationality and phone OS. We collect reach estimates for all combinations of discretized demographic groups for two cities: New York City, USA and Doha, Qatar.

To validate Facebook's inference algorithms for attributes such as country of origin or education level, we look at the correlation of each attribute with 764 indexes from the 2012 to 2016 American Community Survey (ACS)[3] across 176 zip codes in New York. For example, selecting the category "Latinos", a combination of countries of origin in Latin America defined by us, the highest correlated index is the percentage of Latinos, as an ethnicity, in the ACS (c=0.55, pv=0.03).In another example, the lowest correlation with the fraction of users with the only "High School" education attribute was with the index for "a bachelor degree in social science" (c=-0.55, pv=0.03). Overall, we looked at the top 1% absolute correlation values finding significant expected positive and negative correlations for scholarly, country of origin and wealthiness. Regarding age and gender attributes, its distribution differences are subtler across zip codes and could not be verified using this approach. We could not find a survey or census to build a similar comparison for Qatar.

The Interface

The two main components of the visualization, illustrated in Figure 1, are the tree map on the left allowing both the exploration and selection of demographic attributes and the geographic map, color-coded by the share of the population matching selected attributes. As a user clicks on various demographic selections (say, selecting Gender to be Female), all other bars adjust to reflect the demographic characteristics of the selection. Two ways of segmenting geographic areas are possible: New York City visualization demonstrates the pre-defined ZIP codes, which may span only a few blocks at the center of the city, whereas Doha visualization makes use of arbitrary circle selection with the radius of 1km, packed for maximum coverage. All data is loaded upon the webpage load, making the interaction fluid.

Example Use Cases

Census data can be outdated or, in the case of Doha, is not publicly available at all disaggregated by nationality. Facebook advertising data can hence complement existing or missing data to obtain population estimates in near

[1] https://goo.gl/QmNgCy
[2] https://goo.gl/A8WefQ
[3] https://goo.gl/BCKfhB

real-time. For instance, in the Doha figure shown below, Citizenship is selected as Nepal, providing an overview of this population – at around 118 thousand (capturing about a third of the current Qatar-wide estimate[4]) it is overwhelmingly male and young, and using mostly Android phones, which are typically cheaper than iOS devices. Geographically, this group is located in lower left corner of the map, area of the city dubbed as Industrial Area[5], which was built to cater for the Asian expatriate population. Such insights help understand some of the conditions of a cohort, informing decisions about local housing and other public policies. Similarly, a selection of college graduates in New York City presents us with the highlighted areas in Manhattan and nearby neighborhoods, as well as showing a user base predominantly using iOS devices. Monitoring changes in this geographic distribution may help reveal the effects of gentrification or reveal the effects of housing policies.

Privacy Concerns

Our visualization works with aggregate, anonymous data. As such we believe that privacy concerns for our visualization are similar to those of other aggregate population data. However, researchers have pointed out how the Facebook advertising API could in the past be misused to obtain personally identifiable information [5, 4]. Facebook has since closed the corresponding loopholes concerning the audience estimates for so-called "custom audiences"[6]. These types of special audience estimates were never used for this data visualization or any of our related research.

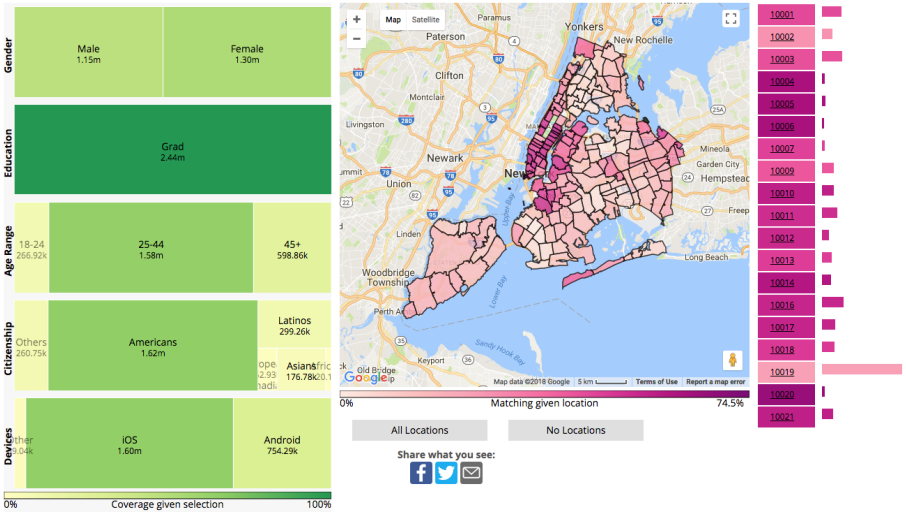
Conclusions

The tool presented illustrates the fine granularity of the demographic data available through the Facebook Marketing API, and shows how such data can be interactively explored. Potential users for this "Facebook population census" include demographers, urban planners, and public policy professionals.

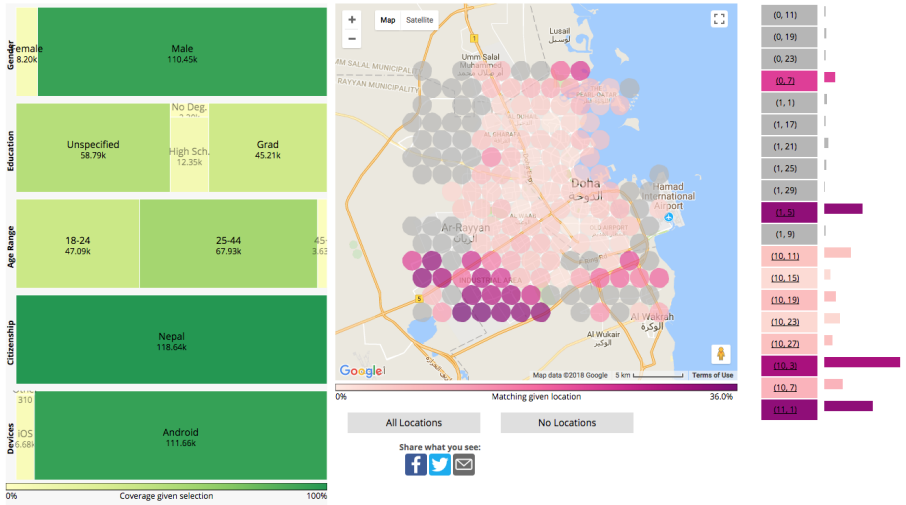
[4] https://goo.gl/DJvKvr
[5] https://goo.gl/Tp7u1Z
[6] See https://goo.gl/SsqK4N for some background on this.

REFERENCES

1. Matheus Lima Diniz Araújo, Yelena Mejova, Michaël Aupetit, and Ingmar Weber. 2017. Visualizing Health Awareness in the Middle East. In *ICWSM*. 725–726.
2. Rumi Chunara and others. 2013. Assessing the Online Social Environment for Surveillance of Obesity Prevalence. *PLOS ONE* 8, 4 (04 2013), 1–8.
3. Justin Cranshaw, Raz Schwartz, Jason I. Hong, and Norman M. Sadeh. 2012. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. In *ICWSM*. 58–65.
4. Aleksandra Korolova. 2010. Privacy Violations Using Microtargeted Ads: A Case Study. In *International Conference on Data Mining Workshops*. 474–482.
5. Giridhari Venkatadri, Athanasios Andreou, Yabing Liu, Alan Mislove, Krishna P. Gummadi, Patrick Loiseau, and Oana Goga. 2018. Privacy Risks with Facebook's PII-based Targeting: Auditing a Data Broker's Advertising Interface. In *SP*. 221–239.
6. Ingmar Weber and Venkata Rama Kiran Garimella. 2014. Visualizing User-Defined, Discriminative Geo-Temporal Twitter Activity. In *ICWSM*. 656–657.
7. Emilio Zagheni, Ingmar Weber, and Krishna Gummadi. 2017. Leveraging Facebook's Advertising Platform to Monitor Stocks of Migrants. *Population and Development Review* 43 (2017), 721–734. Issue 4.



(a) Greater New York City, USA



(b) Doha, Qatar

Figure 1: Audience visualization of New York City, USA with selections of university graduates (a) and Doha, Qatar (b) with Nepali expats (b).