Assessing the Reliability of Facebook's Advertising Data for Use in Demographic Research

André Grow (✉), René Flores, Ilana Ventura, Ingmar Weber, Kiran Garimella, and Emilio Zagheni

Background and Goals

An increasing number of scholars advocate the use of Facebook's advertising platform in demographic research, either for conducting 'digital censuses' that provide information about the compositional characteristics of the population at large, or for recruiting participants for survey research (e.g., Alburez-Gutierrez et al. 2019; Alexander, Polimis, and Zagheni 2019; Cesare et al. 2018; Pötzschke and Braun 2017; Zagheni, Weber, and Gummadi 2017). The efficacy of both approaches depends on the accuracy of the data that the advertising platform provides about Fakebook's users, but so far little is known about the reliability of this data. In this paper, we address this lacuna by comparing Facebook's user classification with users' self-reported information in a demographic survey.

Facebook currently is the largest social media platform, with 2.41 billion monthly users around the world (Facebook 2019). Its business model centers on revenue from online advertising (Zagheni, Weber, and Gummadi 2017), implemented thorough the Ads Manager (AM) platform. The AM enables advertisers to create advertising campaigns that can have various goals, such as creating salience for a given service or product among Facebook users, or generating traffic to an external website. Each campaign can be targeted at specific user groups. These groups can be defined based on several self-reported demographic and personal characteristics (e.g., the user's sex and age) and a set of characteristics that Facebook infers from the user's behavior on the network (e.g., musical interests based on interactions with the Facebook profiles of certain music bands). Prior to launching a campaign, the AM provides an estimate of the expected audience size (i.e., the number of monthly active users who are eligible to be shown a given ad) given the selected combination of user characteristics. This allows advertisers to optimize their definition of target groups (Cesare et al. 2018). Figure 1 provides an example of this, focusing on women age 30–35 years, who live in the United States (US). According to Facebook, there were about 17 million monthly active users who belonged to this group as of September 2019. Once a campaign has started, its advertisements are automatically delivered to the members of the specified user groups, subject to possible competition for advertising space with other advertisers who target the same groups.

Earlier demographic research has used the facilities that the AM provides in one of two ways. A first set of studies have employed the audience estimates that the AM provides prior to launching a campaign for obtaining digital censuses of the user population in a given geographical region. The resulting information was then used to make inferences about the general population. For example, Zagheni, Weber, and Gummadi (2017) used AM-audience estimates to assess the share of foreign born people living in the states of the US, comparing these numbers with data from the 2014 round of the American Community Survey (ACS). Their results showed that the AM-audience estimates were qualitatively similar to the number

✉ André Grow (grow@demogr.mpg.de), Max Planck Institute for Demographic Research, Rostock, Germany

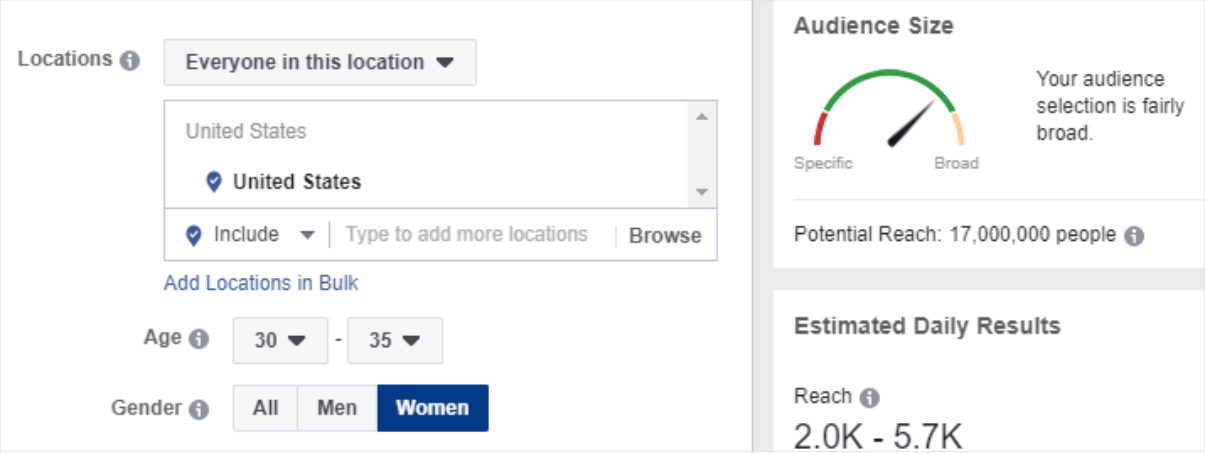

Figure 1. Example of audience-size estimation in the Facebook AM as of September 2019

of migrants observed in the ACS, which suggests that the AM data can be used to study compositional population properties. One central benefit of this approach is that the information that the AM provides is updated continuously and can be collected programmatically through the Facebook Application Programming Interface. This makes it possible to collect population data in a more continuous and more timely manner than often is possible with traditional surveys (e.g., a regular census) or register data.

A second set of studies have used the targeted advertising facilities of the AM to recruit participants for survey research. With this approach, researchers define one or more user groups who should be shown an ad that invites them to participate in an online survey. This ad will then be displayed in the users' Facebook timeline, and when they click on the advertisement, they are directed to an external webpage where they can participate in the survey. Pötzschke and Braun (2017) used this approach for recruiting Polish migrants in several European countries for a survey that queried them about their use of social networking sites, their migration experience, and their socioeconomic background. The authors were able to quickly recruit a large number of participants in a short amount of time, at comparatively low cost. Given Facebook's reach, this approach is particularly attractive when the goal is to recruit members of subpopulations that account only for a very small share of the overall population and that are difficult to identify in existing sampling plans.

As the foregoing illustrates, the Facebook AM has potential for complementing standard instruments in the demographic toolbox. However, it is important to keep in mind that the validity and cost-effectiveness of the two approaches discussed above crucially depend on the accuracy of the user information that Facebook provides. For example, taking a digital census of a population in a given country – possibly broken down by gender, age, and ethnicity – is only feasible if Facebook accurately classifies its users according to these characteristics. While classifying users based on self-reported characteristics such as gender and age may be relatively straightforward, this is likely to be more difficult based on inferred characteristics such as ethnicity. Similarly, the cost-effectiveness of using Facebook for recruiting participants for survey research would be greatly reduced when the generated traffic to the survey would include many users who are not part of the targeted demographic group. Such participants would need to be excluded in the final analysis of the survey and this would

increase the cost per usable questionnaire.

These problems are aggravated by the fact Facebook's category definitions do not follow scientific standards and are often ambiguous (Cesare et al. 2018). To illustrate this, consider how migrants have been identified in earlier research using the AM. Facebook does not include the category 'migrant' in the list of possible user characteristics. Instead, it includes categories that indicate where people lived in the past, such as 'Lived in Colombia (Formerly Expats – Colombia)', which Facebook defines as 'Users who used to live in Colombia who now live abroad'. Earlier research has used these classifications to count the number of migrants in a given country (e.g., Zagheni, Weber, and Gummadi 2017), but it is possible that these categories include people who would not be considered migrants in more traditional data sources.

Taken together, future demographic research that seeks to employ the Facebook AM would benefit from a systematic analysis of the accuracy of Facebook's user classification. In this paper, we will provide precisely such an analysis, focusing on Facebook users in the US.

Study Design

The results that we will report in this paper are part of a larger research project that focuses on the cultural assimilation of Mexican migrants in the US. The goal of this project is to compare immigrants' cultural preferences and political attitudes with that of non-migrants, including pairwise comparisons with native-born non-Hispanic whites, and other members of the US population. To collect the data necessary for this comparison, we have designed an online survey in which respondents are asked detailed questions about their ethnic background, migration history, and other demographic characteristics, as well as their cultural preferences and political attitudes. Recruitment for our survey will take place via the Facebook AM, by means of a stratified advertising campaign.

In more detail, several scholars have pointed out that Facebook's user population tends to differ from the overall population in central demographic characteristics (e.g., Zagheni, Weber, and Gummadi 2017; Zhang et al. 2018). Hence, it is advisable to use a stratified sampling approach to ensure that the participants that are recruited for survey research are representative of the general population (Zhang et al. 2018). With this approach, researchers create one advertising campaign for each population stratum that they are interested in and allocate their advertising funds proportionally to the number of individuals who belong to a given stratum in the overall population. For example, if researchers wanted to ensure that their sample is representative of the US population in terms of its gender (considering the categories men and woman) and age (considering the age groups 20–30, 30–40, and 40–50) composition, they would create one campaign for each possible combination of the two characteristics (leading to 2*3 = 6 separate campaigns), allocating larger advertising funds to larger population strata. Survey values are then re-weighted using post-stratification, potentially combined with multilevel models when issues related to data sparsity arise, in order to make statistical inference from non-representative samples (Wang et al. 2015).

In this paper, we use this approach to indirectly study Facebook's user classification. For this, we employ a feature of the AM that has received little attention in earlier research: each advertising campaign is assigned a unique ID and this ID can be passed on to the target webpage of a given campaign. By recording this ID and matching it with respondents'

answers to the survey, it becomes possible to assess the accuracy of Facebook's user classification. For example, an ad that is targeted at women who are 20–29 years old should only be shown to individuals which Facebook has categorized as belonging to this demographic group. Hence, by recording the ID of the campaign that respondents have clicked on, it becomes possible to infer the categories to which Facebook has allocated them.

Following the work of Zhang et al. (2018), we will focus on US residents and stratify our campaign based on gender (male, female), age (18–24 years, 25–44 years, 45–64 years, 65+ years), and region (Northeast, South, West, Midwest), as well as place of birth (US, Mexico), resulting in a total of 64 non-overlapping strata. The provisional budget allocated to all campaigns together is US$5,000, and we will provide participants with US$5 Amazon gift cards for completing the survey. Figure 2 shows an example of the advertisements that we will use. The field work will start in October 2019 and is scheduled to last for four weeks.

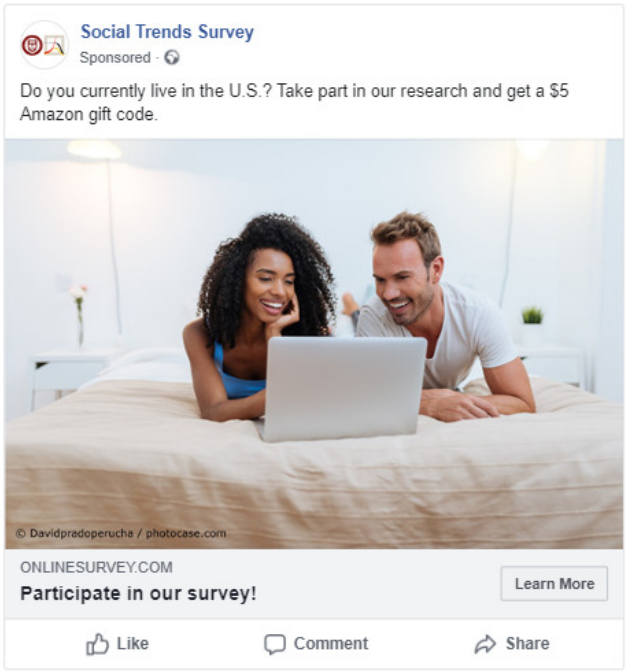

Figure 2. Example of advertisement to be used in our advertising campaign

Expected Findings

Our study is explorative, and we have no hypotheses as to the overall accuracy of Facebook's user classification. However, to the extent that there are inaccuracies, we expect that there will be differences between different types of classification criteria. We expect that classification criteria based on mandatory, user-provided information will be more accurate than classifications based on non-mandatory information that may need to be partially inferred. More specifically, we expect that Facebook's categorization in terms of users' gender and age (based on the user's birth date; mandatory) will show less error than classifications related to their region and place of birth (non-mandatory, partially inferred).

Acknowledgements

We thank Ian Stewart for his contributions to designing the online survey.

References

Alburez-Gutierrez, Diego et al. 2019. "Demography in the Digital Era: New Data Sources for Population Research." In *Book of Short Papers SIS2019*, eds. G. Arbia, S. Peluso, A. Pinna, and G. Rivellini. Pearson, 22–33.

Alexander, Monica, Kivan Polimis, and Emilio Zagheni. 2019. "The Impact of Hurricane Maria on Out-Migration from Puerto Rico: Evidence from Facebook Data." *Population and Development Review* 45(3): 617–30.

Cesare, Nina et al. 2018. "Promises and Pitfalls of Using Digital Traces for Demographic Research." *Demography* 55(5): 1979–99.

Facebook. 2019. “Company-Info.” *Stats*. https://newsroom.fb.com/company-info/ (September 24, 2019).
Pötzschke, Steffen, and Michael Braun. 2017. “Migrant Sampling Using Facebook Advertisements: A Case Study of Polish Migrants in Four European Countries.” *Social Science Computer Review* 35(5): 633–53.
Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. 2015. “Forecasting Elections with Non-Representative Polls.” *International Journal of Forecasting* 31(3): 980–91.
Zagheni, Emilio, Ingmar Weber, and Kirshna Gummadi. 2017. “Leveraging Facebook’s Advertising Platform to Monitor Stocks of Migrants.” *Population and Development Review* 43(3): 721–34.
Zhang, Baobao et al. 2018. “Quota Sampling Using Facebook Advertisements.” *Political Science Research and Methods* (online first).