Leveraging Facebook's Advertising Platform to Monitor Stocks of Migrants

Forthcoming in *Population and Development Review* (Accepted June 2017)

Emilio Zagheni*

Department of Sociology, University of Washington, Seattle

Ingmar Weber

Qatar Computing Research Institute, Doha, Qatar

Krishna Gummadi

Max Planck Institute for Software Systems, Saarbruecken, Germany

August 28, 2017

*corresponding author. E-mail: emilioz@uw.edu; address: 211 Savery Hall, Box 353340, Seattle, WA 98195, USA

Abstract

Online advertising is the main source of revenue for social media companies. Facebook allows advertisers to target users with certain characteristics, such as age, gender, country of origin, education level, or topical interest. Before an ad is launched Facebook's advertising platform provides an estimate of how many users match the provided criteria. This functionality, akin to a 'digital census' over Facebook users, has so far been untapped for demographic research. We show the feasibility of estimating stocks of migrants within and across countries and discuss the potentials and limitations of the data. The lack of much needed information about migrants, together with the rapid global expansion of social media use and the online advertising industry, offer new opportunities to study migration as well as monitor progress towards achieving the goals of the United Nations 2030 Agenda for Sustainable Development.

In the context of development, demographic data are essential to monitor the health and well-being of populations. Measures of mortality, often summarized in terms of life expectancy, are key indicators of health. Measures of fertility and educational attainment of women are related to the well-being of a society and to gender equality. Migration rates could be an indicator of economic vitality or of crisis.

Demographic data are important for monitoring development. Thus the lack of appropriate sources and indicators for measuring progress towards the achievement of targets – like the ones set by United Nations in the context of the "2030 Agenda for Sustainable Development" – is a significant cause of uncertainty. As part of a larger effort to tackle the issue, in 2014 the United Nations asked an independent expert advisory group to make recommendations to bring about a data revolution in sustainable development. Data innovation, like new digital traces from a variety of technologies, is seen as a huge opportunity to inform policy evaluation and to improve estimates and projections.

In this article, we contribute to the development of tools and methods that leverage new data sources for demographic research. More specifically, we present an innovative approach to estimate stocks of migrants using a previously untapped data source: Facebook's advertising platform. This freely available source allows advertisers (and researchers alike) to query information about socio-demographic characteristics of Facebook users, aggregated at various levels of geographic granularity. Overall, This article has three

main goals: 1) to present a new data source that is relevant for demographers; 2) to discuss how demographers can avoid some of the problems related to the analysis of non-representative Web and social media data; 3) to lay out the foundations on which demographers and data scientists can build in the future.

The focus of the paper is on migration, but the general approach that we present could be applied also to demographic indicators related to fertility, health and mortality. Some of the broader issues that we address, like the need for small-area, timely and disaggregated indicators, are key for other disciplines as well. In economics, for instance, quantities like the Gross Domestic Product (GDP) are determined with a lag, and final estimates are produced only after a series of revisions. The need for timely estimates has led to the development of an area of research often referred to as 'nowcasting', or prediction of the present. (Giannone et al., 2008) The rapid expansion of digital traces, such as those that are byproducts of Web search engine queries, have resulted in the development of new approaches for nowcasting (Ginsberg et al., 2009; Choi and Varian, 2012). They also stimulated a reflection about the problem of relying solely on non-representative sources, when in fact the combination of traditional and new data, within a unified framework, would yield improved estimates and predictions. (Lazer et al., 2014).

In the context of migration studies, the lack of timely data about migrants limits our ability to address societal challenges. (Willekens et al., 2016) Improving migration statistics requires timely data, at different levels

of geographic granularity. These statistics would be important for assessing the impact of policy interventions. For example, the effect of policies on migration and the impact of migrant flows on policies can be estimated by leveraging 'natural experiments' when appropriate data are available. Data sets about global migration are important to improve migration theories, to reduce our uncertainty about the present state of migration in the world, and to improve forecasts and population projections.

New statistical approaches to analyze Census data (Abel and Sander, 2014) as well as new methods that leverage innovative Web data sources, like geo-located e-mail logins (Zagheni and Weber, 2012; State et al., 2013), Twitter tweets (Zagheni et al., 2014; Hawelka et al., 2014) and Linkedin profiles (State et al., 2014) have expanded our ability to estimate migration rates and migration flows. The literature in this area has demonstrated the feasibility of using digital breadcrumbs to study migrations, and has provided an assessment of the relative importance of selection bias in determining the overall uncertainty about the estimates.

Although Facebook is the largest social media platform, and a natural choice for the study of migrations, very little work has been done with Facebook data. The main bottleneck has been data access. The existing work in this area is limited to projects carried out by data scientists working at Facebook, and the results are often disseminated via blog posts.[1]

This article is organized as follows. First, we describe the Facebook's Advertising Platform. Second, we present a proof of concept related to esti-

mating stocks of migrants using Facebook data. Third, we evaluate biases in the data and show how accounting for these biases leads to better estimates and predictions. Finally, we discuss the challenges and opportunities ahead in this area of research.

Facebook's Advertising Platform

Facebook's main stream of revenue is online advertising. In an effort to attract advertisers and to improve their return-on-investment, Facebook has developed a targeted advertising platform, called Adverts Manager, that allows advertisers to give detailed specifications on the type of users to whom their ad should be shown [2]. The dimensions that can be targeted include both information explicitly reported by Facebook users, such as their age or gender, as well as information automatically inferred from their interaction on Facebook and affiliate websites, such as their interests. As an illustrative example, Facebook supports showing ads exclusively to Italian expats, aged 18 and over, living in the state of Washington. Before actually launching an ad, which then incurs a cost to the advertiser, Facebook provides an estimate of the selected audience size. In the example above, Facebook reports a "potential reach" of 3,800 users.[3] This reach estimate refers to the number of monthly active users on Facebook that match the described criteria[4]. For our analysis we obtain these reach estimates in a programmatic way via Facebook's Marketing API[5]. As we did not proceed to actually launch an

ad, these data were collected free of charge. We believe that it is important for the demographic community to become aware of this untapped and rich data source that can be leveraged to improve our understanding of population processes in the US as well as in the developing world.

Previous research related to migration has used Facebook Adverts Manager to reach migrants that might not otherwise be included in traditional surveys. More specifically, hard to reach Polish migrants in European countries were sent a Facebook online advertisement that would invite them to participate in a survey. (Pötzschke and Braun, 2016)

Facebook Adverts Manager offers unparalleled opportunities for social sciences, survey research and policy analysis. In this article we focus on the potential for extracting, in a programmatic way and free of charge, demographic characteristics of Facebook users, where Facebook can be thought of as a large census of more than 1.9 billion monthly active users. We believe that combining Facebook data with traditional sources, in order to address issues related to selection bias, is going to generate relevant and new demographic knowledge.

Evaluating migration stocks: A proof of concept

For our analysis, we make use of the category "Expats (*)" that Facebook provides. As an example, for the category "Expats (Mexico)" Facebook

gives an estimate of 8.4 million monthly active users aged 18+ who live in the US. Facebook's Marketing API currently supports 52 countries or territories of origin when targeting expats of a particular origin, such as 'Expats (Mexico)'.[6] Additionally, one can target 'Expats (All)' which also includes users of other countries of origin. In May 2017, Facebook Adverts Manager reports a total of 202 million "expats" 18 years old or older on Facebook, on a global scale. This value is in a range consistent with the United Nations estimate of worldwide foreign-born migrants (244 million people in 2015). Counts of expats in Facebook can be disaggregated at different levels of granularity, including states, cities and metropolitan areas, for the place of residence.

Facebook does not provide a very detailed definition of 'expats'. In the Adverts Manager's documentation expats are defined as "people whose original country of residence is different from the current country". Despite the lack of documentation, we can infer, based on the literature produced by researchers who work internally at Facebook (Herdağdelen et al., 2016), that there are two factors that play a key role in the estimation of expats at Facebook. The first one is the self-reported "current city" and "hometown" in the list of "places you have lived" that people fill in for their Facebook profile. The second one is the network structure of friendships (e.g., having at least two Facebook friends in the home country and two Facebook friends in the destination country). Herdağdelen et al. (2016), working internally at Facebook, generated estimates of migrants based on the set of variables

that we just described. In their publication they provided the ranking of the top 11 immigrant communities in the US on Facebook, according to their country of origin (Mexico, India, Philippines, Puerto Rico, El Salvador, Dominican Republic, Guatemala, Canada, Honduras, Cuba and Colombia). This ranking is almost exactly identical to the one provided by Facebook Adverts Manager. The only difference is that, based on Adverts Manager, Vietnam replaces Canada in the list of top countries of origin in the US. However, the difference between the two countries is minimal (0.03% of the US population of Facebook users is estimated to come from Vietnam, whereas 0.026% is estimated to come from Canada). Although Facebook could estimate "expats" in many different ways, including very complex algorithms, we are confident that 'hometown', 'current city' and the network structure of Facebook friendships are among the key components of their estimation process.

Here we report two illustrative examples that serve as proof of concept for the use of the Facebook data set for demographic research and, in particular, for the study of migrations. Figure 1 shows the relationship between Facebook estimates of fraction of expats in US states (among Facebook users, by country of origin, as of 2016) and the fraction of foreign born people from the latest available data from the American Community Survey (ACS, 2014).[7] The dashed line is a 45 degree line. The figure shows that Facebook's estimates are highly correlated with the ones from the ACS. For example, a basic linear model fits the data extremely well: about 94% of the variability

in the data is explained by a straight line through the observations. Similarly, transforming the data in the log scale to address the skewed distribution of the rates (see inset) shows that the relationship is very linear. Figure 1 also highlights that there are systematic biases: as most data points lie above the 45-degree line, Facebook tends to underestimate migration stocks across US states.

While the previous example was for the US context, Figure 2 expands the scope to other countries and shows the relationship between stocks of migrants from the Facebook data set (2016), and the respective estimates from the World Bank. The data points are for the 96 countries where there are at least 1 million monthly active Facebook users. There is a relatively strong correlation between estimates of international stocks of migrants in Facebook and the ones obtained by the World Bank. On a log scale, Facebook data explain approximately 60% of the variability in the World Bank estimates. Facebook can be thought of as a biased census, with biases that differ across continents as well as across socio-economic strata of the society. In the context of the analysis for Figure 1, the bias is expected to be quite small and not so heterogeneous, since we are looking at fractions of migrants across states of the same destination country, the US. In other words, we might expect that, say, Mexicans in the US who use Facebook are not necessarily representative of the underlying population of Mexicans in the US. However, we would not expect that the level of bias for Mexicans who live in Texas and use Facebook is substantially different from the one of Mexicans

who use Facebook and live in, say, California. It is important to notice that the bias is also relatively small when we compare immigration rates across countries, as shown in Figure 2. This indicates that, despite measurement issues and selection bias, it is potentially feasible to derive robust estimates of demographic indicators from tabulations of Facebook users.

Understanding the bias in the data

In order to improve our understanding of the direction and size of the biases in our data set, we disaggregated our estimates by age and sex. Figure 3 shows profiles of stocks of migrants by age and sex for Mexicans in California and in Texas, two of the states with the largest populations of Mexicans in the US. The plots compare estimates from the American Community Survey with those obtained from Facebook Adverts Manager.

We observe that for men and women, and for both states, the pattern of the bias across age groups is very similar. Assuming that the American Community Survey provides an unbiased picture of the stocks of migrants, we notice that Facebook data closely match, or slightly overestimate, stocks of migrants for young age groups (people in the early 20s or younger). For older age groups, Facebook data consistently underestimate the fraction of migrants. The underestimation problem could be related to a number of reasons. Facebook users are not representative of the underlying population of Mexicans. Users in different age groups may be selected for different

characteristics. Also, older Mexicans may spend less time on Facebook than younger people or may report less information about their home country. Therefore it may be harder for Facebook to estimate that their country of origin is Mexico. In addition there are measurement issues: Facebook's definition of expats is not necessarily equivalent to foreign-born; people may misreport their hometown, and the data from the American Community Survey is not exactly for the same time period as the one from Facebook. There may be many reasons behind the discrepancy between data from the American Community Survey and data from Facebook. However, for our purposes it is important to note that, regardless of the factors that generate the discrepancy, there are similarities in the patterns across states. These patterns can be leveraged to borrow strength in the production of estimates across states. In other words evaluating the pattern of bias for Mexicans in California is useful to improve predictions of Mexicans in Texas.

It is also important to note that there may be factors related to bias and measurement error that are specific to the country of origin. For example, Figure 4 shows profiles of stocks of migrants by age and sex for Filipinos and Filipinas in California and Texas. We observe that the pattern of bias for Filipinas is qualitatively similar to the one of Mexican women, with estimates from Facebook closely matching the ones from ACS for young age groups, and underestimating at older ages. However, for Filipinos we observe that Facebook data overestimate migration at older ages. The trend is consistent across the two US states, indicating that there may be some factors specific

to migrants from the Philippines.

In order to evaluate the bias in the data in a more systematic way, we estimated the parameters of the following linear regression model:

$$\begin{aligned} log(\text{ACS foreign born pop}_{ij}^{z}) = \beta_0 + \beta_1 log(\text{Facebook expats}_{ij}^{z}) + \\ + \beta_2 \mathbb{1}(\text{Origin 1}) + \cdots + \beta_{30} \mathbb{1}(\text{Origin 29}) + \\ + \beta_{31} \mathbb{1}(\text{Age group 1}) + \cdots + \beta_{38} \mathbb{1}(\text{Age group 8}) + \\ + \epsilon_{ij}^{z} \end{aligned} \tag{1}$$

where (ACS foreign born pop_{ij}^{z}) is the number of people in the age-sex group z born in country i and living in US state j. (Facebook expats_{ij}^{z}) is the number of expats in the age-sex group z from country i who live in US state j. $\mathbb{1}$(Origin 1) is an indicator variable for country of origin 1. $\mathbb{1}$(Age group 1) is an indicator variable for age group 1. In this model, each country of origin has its own coefficient, which serves as a 'level' parameter. Each age group also has its own coefficient that serves as a 'shape' parameter.

Table 1 shows the results for the regression model described in equation 1. In addition to each country of origin having its own parameter, we observe that young age groups have negative coefficients, whereas older age groups have positive coefficients. This indicates that, on average, Facebook data overestimate migration stocks for younger age groups and underestimate the stocks for older age groups.

Understanding bias can be used for predictive purposes. For example, consider the situation in which we would like to estimate the number of foreign born people from country i living in US state j. We can consider a regression model where aggregate estimates of foreign born people from the ACS are regressed against aggregate estimates of Facebook expats, without accounting for biases that may vary by age and country of origin (we refer to this model as the *naive* model). Alternatively, we can use the model described in equation 1 (we refer to this model as the *age-origin* model).

In order to test the predictive capacity of the two models we split our sample in two, a training set and a test set. The training set includes data for 40 randomly selected US states (80% of the 50 states in the US). The model parameters are estimated using the training data set. Then we make predictions about the quantity of interest, the total number of people born in country i and living in US state j, for the states in the test data set. We measure the predictive capacity of the models by calculating the Mean Absolute Percentage Error (MAPE) for the out of sample prediction on the test data set. We record the average MAPE from 50 different trials in which each time a new training set (of 40 states) and a the respective test set (with the remaining 10 states) are generated.

We obtained that, on average, the MAPE for the *naive* model is equal to 56%. The MAPE for *age-origin* model is significantly lower, equal to 37%. This suggests that, by disaggregating by age and country of origin, we can learn about patterns of bias and borrow strength across states. This helps

us improve the predictive capacity of our models. The result indicates that, although a number of sources of bias exist, there are also systematic patterns in the bias that can be estimated and leveraged.

The Road Ahead: Challenges and Opportunities

As Internet penetration rates in the poorest countries of the world are likely to increase at a faster rate than the development of mature registration systems, developing statistical tools that combine traditional and new sources of information is likely an effective approach to satisfy the demand for monitoring demographic rates.

We showed that in Facebook data there is signal to produce timely estimates of migration, and we presented a data resource that is available to the research community, but that has not been leveraged for demographic research yet. The data, that can be obtained via the Facebook Advertising Platform, offer an unprecedented breadth of dimensions including educational attainment, job sector, life events such as birth of children, along with many more. All these variables are regularly updated and can be downloaded free of charge, within certain rate limits. In other words, the data set can be thought of as a census that is continually updated. We offered examples related to the study of migration, but we expect that these data will be used also for other types of demographic research.

These data, when used to complement existing statistics, can contribute to the mission of the United Nations within the framework of the so-called 'data revolution'. However, to fully harness the value of these data, it is important to be aware of the limitations and challenges related to working with several types of potential biases or sources of uncertainty.

One key limitation of using Facebook advertising data is that the variables available are not necessarily documented according to standards for scientific research. For example, it is not completely clear how Facebook produces estimates for categories like "Expats (Mexico)." We showed that existing scientific articles written by research scientists at Facebook give us some information about the general approach that Facebook uses to generate estimates. However, the lack of detailed documentation introduces measurement errors that are hard to disentangle from biases related to selection and non-representativeness of the users, or other noise and inconsistencies in the data. In addition, as Facebook improves its algorithms to estimate quantities like 'expats', we may observe discontinuities in the time series, similarly to what we observed for the index of Google queries provided by Google Trends. This means that models that rely on these data may have to be re-calibrated on a regular basis.

These data issues raise serious challenges for producing robust estimates. However, we believe that statistical demographers are well-equipped to deal with noisy and imperfect data, and to develop methods that account for data quality. Combining traditional and new data sources is key to make

progress in this area. In the context of modeling migration flows, demographers have developed a number of creative techniques to deal with inadequate data (Rogers et al., 2003; De Beer et al., 2010; Raymer et al., 2013; Wiśniowski et al., 2016; Zagheni and Weber, 2015). Bayesian hierarchical models have proven particularly useful for dealing with issues related to harmonization and pooling of information across space and time.

In the context of migration stocks, research questions and issues are different from those related to migration flows. However, when the aim is to develop tools for small area estimation, or to combine data, or to borrow strength from countries with better quality data, some of the same ideas that have been developed for the statistical analysis of migration flows can inform the development of methods for the study of stocks. The overall goal remains the development of statistical models to incorporate information from a number of data sources and to evaluate biases and uncertainty when estimating and predicting.

In this article, we focused on the use of Facebook data for estimating migration. Although the scope is global, we concentrated on examples for the United States as a country of destination. For the United States we have good traditional data on stocks of migrants, that can be used to evaluate the potentials of the data and approaches. The main advantage of complementing the ACS with Facebook data is that it would enable researchers to address the stochasticity of small area estimates, which can be quite noisy, given the relatively small sample size and the level of disaggregation needed

(e.g., number of migrants from country X living in county Y, by age, and sex). In developing countries, where there is not a survey continuously run like the ACS, the impact of combining Facebook data and traditional data within a solid statistical framework would be bigger. In this paper, we showed that there is signal in the data from Facebook and that leveraging this 'digital census' shows promise. The next developments would build on this first step.

Facebook Adverts Manager is relevant not only to obtain estimates of demographic characteristics of Facebook users. The tool has a broader relevance for the research community, and in particular survey research. Once a specific group of Facebook users is identified, advertisers can target it with ads, upon paying a fee. Similarly, researchers can use Facebook as a sampling frame to target specific populations with 'ads' intended to recruit them for a specific survey. In the context of hard-to-reach populations and in developing countries, this feature will offer new and exciting opportunities for survey researchers to get the pulse of populations. The combination of a very large sampling framework together with the development of techniques that leverage post-stratification to extract information from non-representative samples (see, for instance, Wang et al., 2015) will offer new opportunities for creative survey research.

The scientific community will face not only technical questions, but also ethical ones. While digital advertising can be 're-purposed' to empower researchers, the same tools might also be exploited to perpetuate various forms of discrimination or to identify vulnerable populations. For example, the tar-

geting tool could be used to provide differential information to different racial groups. Recently, a ProPublica investigation showed that it is potentially possible to target Facebook users who are house hunting while also excluding anyone with an affinity for groups like African-American, Asian-American or Hispanic.[8] Although Facebook policies prohibit the use of targeting options to discriminate, the tool is very powerful and could be misused. Similarly, Facebook Adverts Manager could potentially be used to identify geographic areas with rapid influxes of vulnerable populations, like refugees. In the context of war zones, this could expose populations of migrants to risks. As the Facebook tool expands horizons for discovery, a broad discussion about principles for research ethics, privacy protection and responsible conduct needs to reflect the new technological landscape.

Acknowledgments

We would like to thank Matheus Araujo and Kivan Polimis for assistance with data collection. We also would like to thank two anonymous reviewers, the participants to the PAA 2017 session on Big Data analysis in Demography and the participants to the IUSSP/ICWSM17 workshop on Social Media and Demographic Research for valuable feedback.

References

Abel, G. J. and Sander, N. (2014). Quantifying global international migration flows. *Science*, 343(6178):1520–1522.

Choi, H. and Varian, H. (2012). Predicting the present with google trends. *Economic Record*, 88(s1):2–9.

De Beer, J., Raymer, J., Van der Erf, R., and Van Wissen, L. (2010). Overcoming the problems of inconsistent international migration data: A new method applied to flows in europe. *European Journal of Population/Revue européenne de Démographie*, 26(4):459–481.

Giannone, D., Reichlin, L., and Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4):665–676.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., and Ratti, P. K. . C. (2014). Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41:260–271.

Herdağdelen, A., Adamic, L., Mason, W., et al. (2016). The social ties of

immigrant communities in the united states. In *Proceedings of the 8th ACM Conference on Web Science*, pages 78–84. ACM.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of google flu: traps in big data analysis. *Science*, 343(6176):1203–1205.

Pötzschke, S. and Braun, M. (2016). Migrant sampling using facebook advertisements: A case study of polish migrants in four european countries. *Social Science Computer Review*.

Raymer, J., Wiśniowski, A., Forster, J. J., Smith, P. W., and Bijak, J. (2013). Integrated modeling of european migration. *Journal of the American Statistical Association*, 108(503):801–819.

Rogers, A., Raymer, J., and Willekens, F. (2003). Imposing age and spatial structures on inadequate migration-flow datasets. *The Professional Geographer*, 55(1):56–69.

State, B., Rodriguez, M., Helbing, D., and Zagheni, E. (2014). Migration of professionals to the U.S. - evidence from linkedin data. In *Social Informatics*, pages 531–543.

State, B., Weber, I., and Zagheni, E. (2013). Studying inter-national mobility through IP geolocation. In *ACM International Conference on Web Search and Data Mining*, pages 265–274.

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015). Forecasting elec-

tions with non-representative polls. *International Journal of Forecasting*, 31(3):980–991.

Willekens, F., Massey, D., Raymer, J., and Beauchemin, C. (2016). International migration under the microscope. *Science*, 352(6288):897–899.

Wiśniowski, A., Forster, J. J., Smith, P. W., Bijak, J., and Raymer, J. (2016). Integrated modelling of age and sex patterns of european migration. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Zagheni, E., Garimella, V. R. K., Weber, I., and State, B. (2014). Inferring international and internal migration patterns from twitter data. In *International World Wide Web Conference*, pages 439–444.

Zagheni, E. and Weber, I. (2012). You are where you e-mail: using e-mail data to estimate international migration rates. In *Web Science*, pages 348–351.

Zagheni, E. and Weber, I. (2015). Demographic research with non-representative internet data. *International Journal of Manpower*, 36(1):13–25.

Notes

[1]See, for instance, https://www.facebook.com/notes/facebook-data-science/coordinated-migration/10151930946453859/

[2]This platform can be accessed for free, by anybody who has a Facebook account, at https://www.facebook.com/ads/manager/creation/creation/.

[3]As of May 2017.

[4]https://www.facebook.com/business/help/624074880953806

[5]https://developers.facebook.com/docs/marketing-apis

[6]Here is the list of the 52 supported countries and territories of origin that we considered for our analysis: Argentina, Australia, Austria, Bangladesh, Brazil, Cameroon, Canada, Chile, China, Colombia, Egypt, Ethiopia, France, Germany, Greece, Hong Kong, Hungary, India, Indonesia, Ireland, Israel, Italy, Japan, Kenya, Malaysia, Mexico, Morocco, Nepal, New Zealand, Nigeria, Pakistan, Peru, Philippines, Poland, Portugal, Puerto Rico, Russia, Romania, Saudi Arabia, Senegal, Serbia, Singapore, Spain, South Korea, Switzerland, South Africa, Turkey, United Arab Emirates, United Kingdom, United States of America, Venezuela, and Vietnam.

[7]The countries or territories of origin included in Figure 1 are: India, Philippines, Spain, Turkey, France, Germany, Poland, Italy, Ireland, Hungary, Canada, China, Puerto Rico, Indonesia, UK, Australia, Portugal, Nepal, UAE, Singapore, Austria, Greece, Japan, Mexico, Israel, Russia, Saudi Arabia, Malaysia, Romania, South Korea, Vietnam.

[8]See, for instance, https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race

Figures

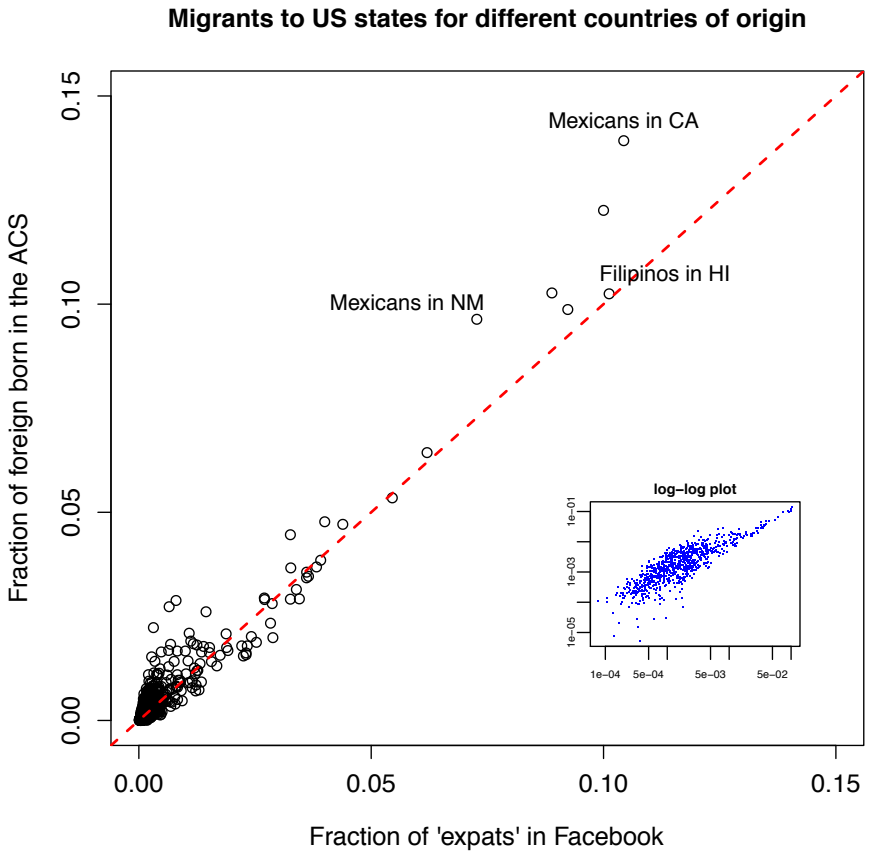

Figure 1: Relationship between Facebook estimates of fraction of expats in US states (2016), by country of origin, and the respective estimates from the American Community Survey (ACS 2014). The red dashed line is a 45 degree line. The inset shows the same observations on a log-log scale. Note: the plot includes state-country pairs where the number of Facebook expats is bigger than 1,000.

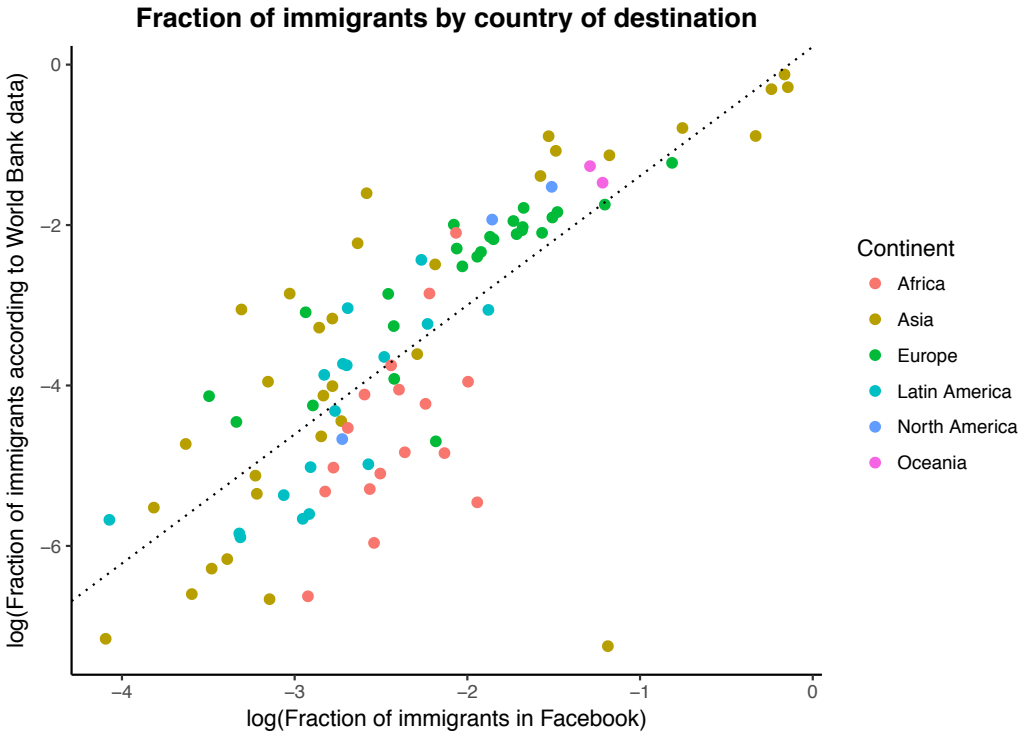

Figure 2: Relationship between stocks of migrants from the Facebook data set, for countries with at least one million Facebook users as of 2016, and the respective estimates from the World Bank (2015). The data points indicate the fraction of immigrants in the population, on a log scale, by country of destination, color-coded by continent. The dashed line is the OLS regression line through the data.

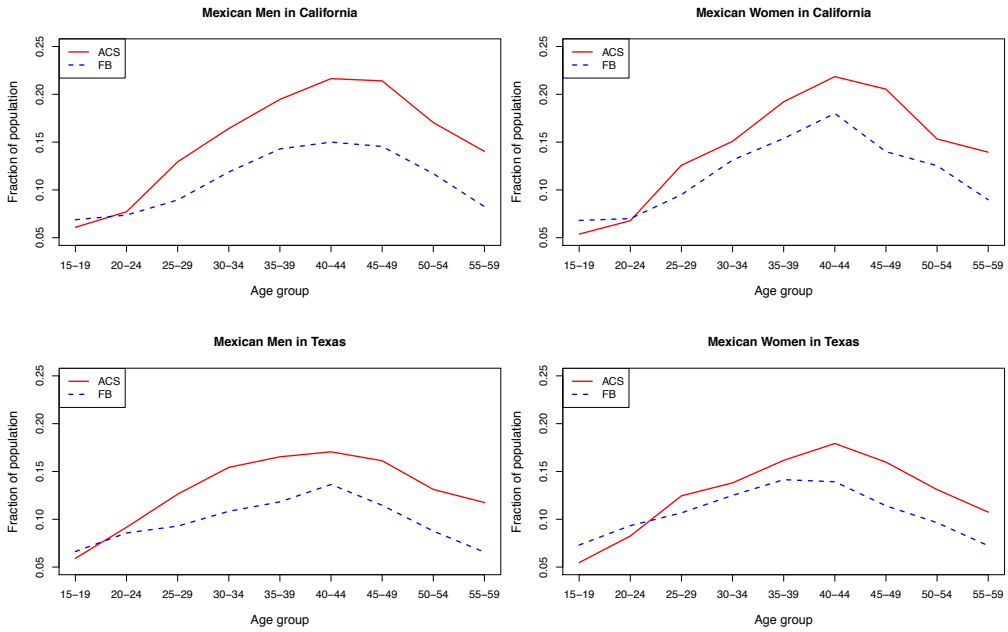

Figure 3: Profiles of stocks of migrants by age and sex for Mexicans in California and in Texas (2016). The solid red line indicates estimates from the American Community Survey (2014). The dashed blue line indicates estimates from Facebook Adverts .

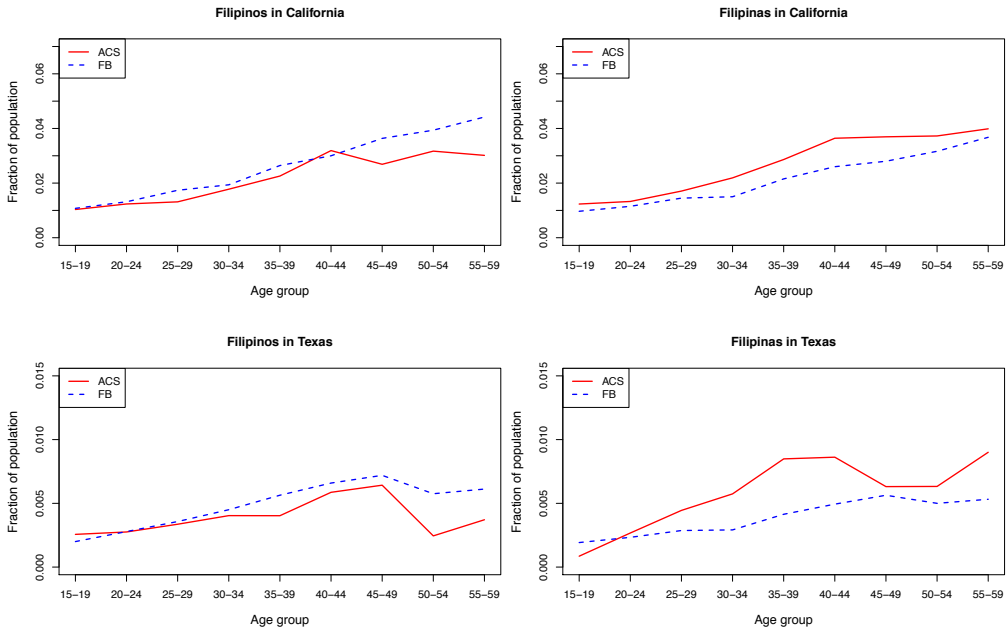

Figure 4: Profiles of stocks of migrants by age and sex for Filipinos and Filipinas in California and Texas (2016). The solid red line indicates estimates from the American Community Survey. The dashed blue line indicates estimates from Facebook Adverts .

Tables

Table 1: Summary of results for the regression model described in equation 1

	log(Foreign-born Population - ACS)
log(FB Expats Population)	0.744*** (0.005)
Austria	0.420*** (0.083)
Canada	0.200*** (0.051)
China	1.132*** (0.050)
France	0.013 (0.056)
Germany	0.879*** (0.050)
Greece	1.443*** (0.064)
Hungary	0.264*** (0.078)
India	0.648*** (0.051)
Indonesia	−0.223*** (0.065)
Ireland	0.193** (0.064)
Israel	0.077 (0.063)
Italy	0.051 (0.057)
Japan	0.538*** (0.052)
Malaysia	0.159* (0.068)
Mexico	0.540*** (0.052)
Nepal	−0.018 (0.062)
Philippines	0.098 (0.051)
Poland	0.526*** (0.060)
Portugal	0.479*** (0.067)
Puerto Rico	0.136* (0.053)
Romania	0.174** (0.059)
Russia	1.069*** (0.052)
Singapore	0.367*** (0.075)
South Korea	0.811*** (0.051)
Spain	0.041 (0.060)
Turkey	0.044 (0.060)
UAE	0.376*** (0.099)
UK	−0.634*** (0.055)
Vietnam	0.301*** (0.052)
Age group (20-24)	−0.483*** (0.032)
Age group (25-29)	−0.291*** (0.032)
Age group (30-34)	−0.010 (0.031)
Age group (35-39)	0.094** (0.031)
Age group (40-44)	0.301*** (0.031)
Age group (45-49)	0.309*** (0.031)
Age group (50-54)	0.460*** (0.031)
Age group (55-59)	0.519*** (0.031)
Constant	1.374*** (0.052)
N	13,328
R^2	0.698
Adjusted R^2	0.697
Residual Std. Error	0.813 (df = 13289)
F Statistic	807.060*** (df = 38; 13289)

*p < .05; **p < .01; ***p < .001