Leveraging Facebook's Advertising Platform to Monitor Stocks of Migrants

Emilio Zagheni
Ingmar Weber
Krishna Gummadi

Given the importance of demographic data for monitoring development, the lack of appropriate sources and indicators for measuring progress toward the achievement of targets—like the United Nations' "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.[1] Data innovation, like new digital traces from a variety of technologies, is seen as a significant 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. 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 to query information about socio-demographic characteristics of Facebook users, aggregated at various levels of geographic granularity. We have three main goals: i) to present a new data source that is relevant for demographers; ii) to discuss how demographers can avoid some of the problems related to the analysis of non-representative Web and social media data; and iii) to lay the foundations on which demographers and data scientists can build in the future.

Our focus is on migration, but 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, Reichlin, and Small 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). These approaches also stimulated 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 more accurate 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. Datasets about global migration are integral to improving migration theories, to reducing uncertainty about the present state of migration in the world, and to refining forecasts and population projections.

New statistical approaches to analyze census data (Abel and Sander 2014) and methods that leverage innovative Web data sources, like geolocated e-mail logins (Zagheni and Weber 2012; State, Weber, and Zagheni 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 flows. Research 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. Work in this area is limited to projects carried out by data scientists working at Facebook itself, and the results are often disseminated via blog posts.[2]

This article is organized as follows. First, we describe Facebook's advertising platform. Second, we present a proof of concept related to estimating 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 advertisers' return-on-investment, Facebook has developed a targeted advertising platform, called Adverts Manager, that allows advertisers to give detailed specifications of the type

of users to whom their ad should be shown.[3] The dimensions that can be targeted include information explicitly reported by Facebook users, such as their age or sex, and information automatically inferred from their interaction on Facebook and affiliate websites, such as their interests. As an 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.[4] This reach estimate refers to the number of monthly active users on Facebook who match the described criteria.[5] For our analysis, we obtain these reach estimates via Facebook's Marketing API.[6] Because we did not proceed to launch an ad, these data were collected free of charge.

Facebook Adverts Manager offers unparalleled opportunities for social sciences, survey research, and policy analysis. Previous research related to migration has used Facebook Adverts Manager to reach migrants who 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 invited them to participate in a survey (Pötzschke and Braun 2016). Here 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, will generate relevant and new demographic knowledge.

Evaluating migration stocks: A proof of concept

We make use of Facebook's category "Expats (*)."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.[7] Additionally, one can target "Expats (All)," which also includes users of other countries of origin. In May 2017, Facebook Adverts Manager reported a global total of 202 million expats 18 years old or older on Facebook. This number is in a range consistent with the United Nations estimate of 244 million worldwide foreign-born migrants in 2015. Counts of expats on Facebook can be disaggregated at different levels of residential granularity, including states, cities, and metropolitan areas.

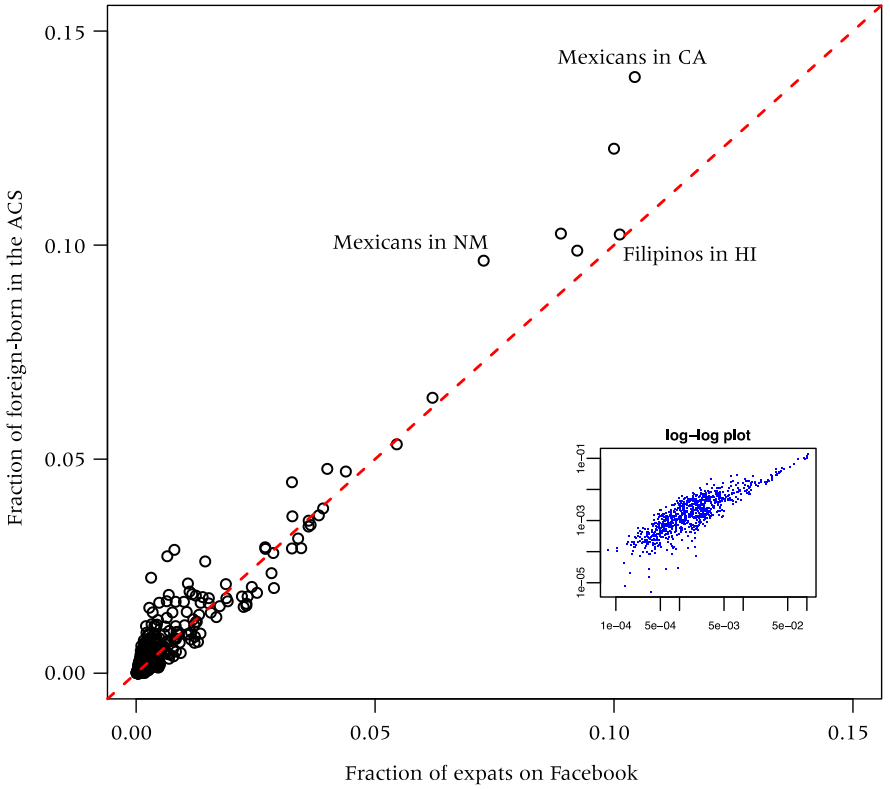
Facebook does not provide a detailed definition of expats. In the Adverts Manager's documentation, expats are defined as "people whose original country of residence is different from the current country." Despite the lack of documentation, we can infer, based on the literature produced by researchers who work internally at Facebook (Herdağdelen et al. 2016), that

two factors play a key role in Facebook's estimation of expats. The first 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 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 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 percent of the US population of Facebook users is estimated to come from Vietnam, 0.026 percent from Canada). Although Facebook could estimate expats in many different ways, including using 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 dataset for demographic research and, in particular, for the study of migrations. Figure 1 shows the relationship between Facebook estimates of the 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 2014 American Community Survey (American Community Survey 2014).[8] Facebook's estimates are highly correlated with those from the American Community Survey (ACS). For example, a basic linear model fits the data extremely well: about 94 percent of the variability in the data is explained by a straight line through the observations. Similarly, transforming the data in the log scale (see inset) to account for the skewed distribution of the rates shows that the relationship is very linear. Figure 1 also indicates that there are systematic biases: since 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, Figure 2 expands the scope to other countries and shows the relationship between the fraction of expats in the Facebook dataset (2016) and the respective estimates of foreign-born population from the World Bank (2015). The data points are for the 96 countries with 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 percent of the variability in the World Bank estimates. Facebook can be thought of as a

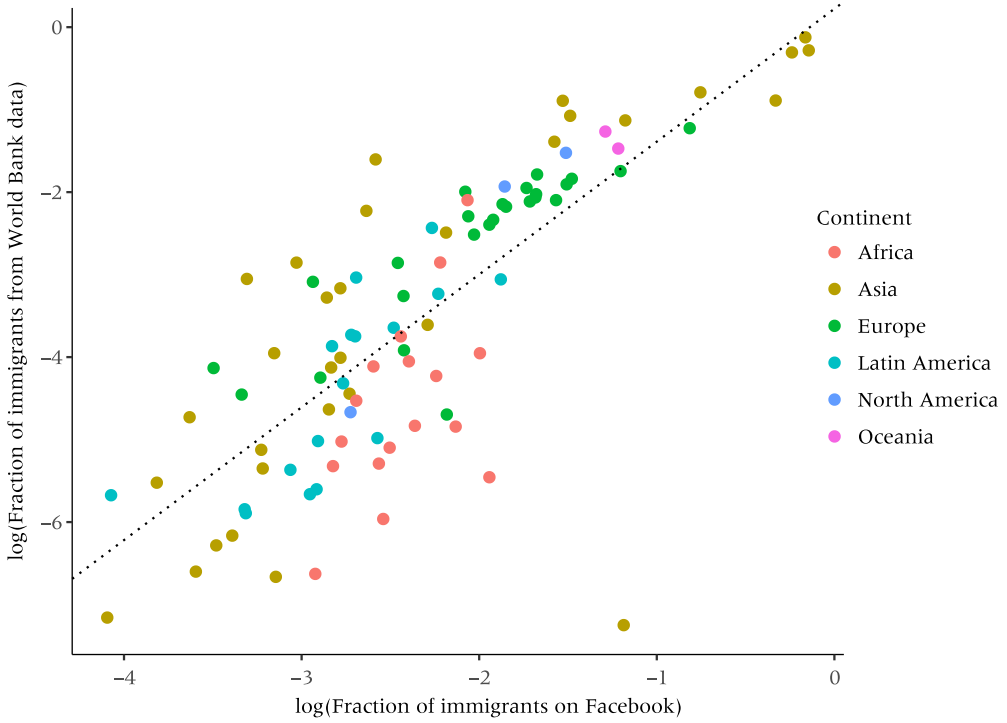
FIGURE 1 Relationship between Facebook estimates of the fraction of expats in US states (2016), by country of origin, and the respective estimates from the 2014 American Community Survey



NOTES: The plot includes state/country pairs where the number of Facebook expats exceeds 1,000. The inset shows the same observations on a log-log scale. The dashed line is a 45-degree line.

biased census, with biases that differ across continents as well as across socioeconomic strata of a given country. In the context of the analysis for Figure 1, the bias is expected to be quite small and not very 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 overall 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 level for Mexicans who use Facebook and live in 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.

FIGURE 2 Relationship between stocks of migrants from the Facebook dataset, for countries with at least one million Facebook users as of 2016, and the respective estimates from World Bank (2015)



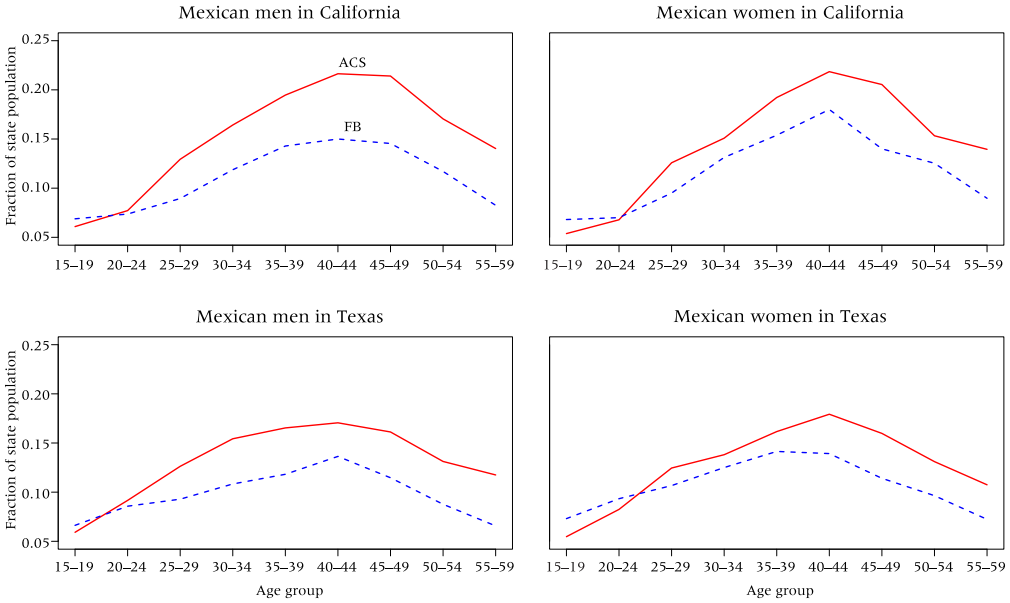
NOTES: The data points indicate the fraction of migrants in the population, on a log scale, by country and continent of destination, color-coded by continent. The dotted line is the OLS regression line through the data.

Understanding bias in the data

To improve our understanding of the direction and size of the biases in our dataset, 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.

For men and women, and for both states, the pattern of bias across age groups is very similar. Assuming that the ACS provides an unbiased picture of the stocks of migrants, 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 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

FIGURE 3 Facebook and ACS profiles of stocks of migrants by age and sex for Mexicans in California and in Texas

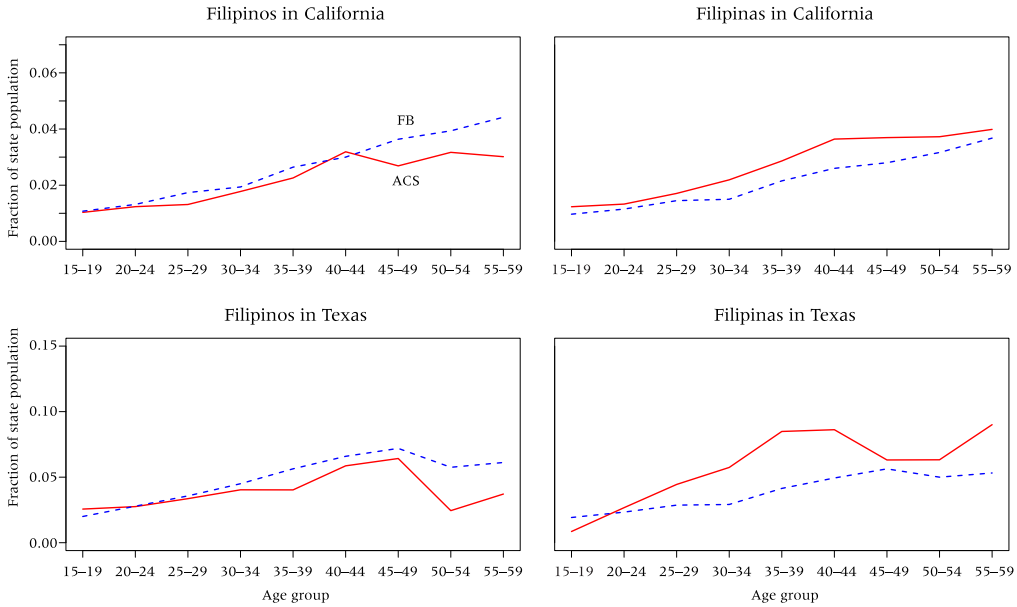


SOURCES: American Community Survey (ACS 2014); Facebook Adverts Manager.

characteristics. Also, older Mexicans may spend less time on Facebook than younger people or may report less information about their home country, making it 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 ACS are not exactly for the same time period as the data from Facebook. There may be many reasons behind the discrepancy between data from the ACS and data from Facebook. For our purposes, regardless of the factors that generate the discrepancy, there are similarities in the patterns across states. These patterns can be leveraged to increase the validity of estimates across states. In other words, evaluating the pattern of bias for Mexicans in California is useful for improving predictions of Mexicans in Texas.

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. The pattern of bias for Filipinas is qualitatively similar to that for Mexican women, with estimates of migration from Facebook closely matching those from the ACS for young age groups and underestimating at older ages. For Filipinos, however, Facebook data overestimate migration at older ages. The trend is consistent across the two US

FIGURE 4 Facebook and ACS profiles of stocks of migrants by age and sex for Filipinos and Filipinas in California and in Texas



SOURCES: American Community Survey (ACS 2014); Facebook Adverts Manager.

states, indicating that there may be some factors specific to migrants from the Philippines.

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

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

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. I(Origin 1) is an indicator variable for country of origin 1. I(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, which serves as a "shape" parameter.

Appendix Table 1 shows the results for the regression model described in equation (1). In addition to each country of origin having its own parameter, 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 stocks for older age groups.

Understanding bias can be useful for predictive purposes. Assume, for example, that we wish 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) (which we refer to as the *age-origin* model).

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 percent of the 50 states). The model parameters are estimated using the training dataset. We then make predictions about the quantity of interest, the total number of people born in country *i* and living in state *j*, for the states in the test dataset. 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 dataset. We record the average MAPE from 50 trials in which each time a new training set (of 40 states) and the respective test set (with the remaining 10 states) are generated.

The MAPE for the naive model, on average, is 56 percent. The MAPE for the age-origin model is significantly lower at 37 percent. This suggests that disaggregating by age and country of origin illuminates patterns of bias, increases the validity of estimates across states, and improves the predictive capacity of our models. In summary, 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

Because internet penetration rates in the world's poorest countries are likely to increase at a faster rate than the creation of mature registration systems, developing statistical tools that combine traditional and new sources of information is likely to be an effective approach to monitoring demographic rates.

We showed that Facebook offers a data resource that is available to the research community but that has not yet been leveraged for demographic research. The data, which can be obtained via Facebook's Adverts Manager, offer an unprecedented breadth of dimensions, including educational

attainment, job sector, life events such as birth of children, and many more. All of these variables are regularly updated and can be downloaded free of charge within certain rate limits. In other words, the dataset can be thought of as a continually updated census. 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 data revolution. However, to fully harness the value of these data, one must be aware of the limitations and challenges related to working with several potential biases or sources of uncertainty.

One key limitation of using Facebook advertising data is that the variables 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 (*). We showed that scientific articles written by researchers at Facebook give us some information about the approach used 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, similar 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.

While these data issues raise serious challenges for producing robust estimates, 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 making progress in this area. In the context of modeling migration flows, demographers have developed techniques to deal with inadequate data (Rogers, Raymer, and Willekens 2003; De Beer et al. 2010; Raymer et al. 2013; Wiśniowski et al. 2016; Zagheni and Weber 2015). Bayesian hierarchical models are 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 increase the validity of estimates by including 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.

We focused on the use of Facebook data for estimating migration, concentrating 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 to 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, which lack surveys like the ACS, the payoff resulting from combining Facebook data and traditional data within a solid statistical framework would be greater. We demonstrated the promise of leveraging the digital census contained in Facebook data. The next developments would build on this first step.

Facebook's Adverts Manager is relevant not only for estimating demographic characteristics of Facebook users. The tool has a broader relevance for the research community, in particular for survey research. Once a specific group of Facebook users is identified, advertisers can target it with ads after paying a fee. Similarly, researchers can use Facebook as a sampling frame to target specific populations with ads intended to recruit them for a survey. In the context of hard-to-reach populations and in developing countries, this feature will offer fruitful opportunities for survey researchers to take the pulse of populations. The combination of a very large sampling framework and the development of techniques that use sampling weights to adjust for under- or over-representation of certain groups in the population (e.g., post-stratification) in order to extract information from non-representative samples (Wang et al. 2015) will offer new opportunities for creative survey research.

The scientific community will face both technical and ethical questions. 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 and target vulnerable populations. For example, the targeting tool could be used to provide differential information to different racial groups. Recently, a ProPublica investigation showed that it might be possible to target Facebook users who are house hunting while also excluding anyone with an affinity for groups like African-Americans, Asian-Americans, or Hispanics.[9] Although Facebook policies prohibit the use of targeting options to discriminate, the tool is very powerful and could be misused. Similarly, Facebook's Adverts Manager could potentially be used to identify geographic areas with rapid influxes of vulnerable populations, such as refugees. In the context of war zones, this could expose 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.

APPENDIX TABLE 1 Summary of results for the regression model described in Equation 1

	log(ACS estimate of foreign-born population)
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

Notes

We thank Matheus Araujo and Kivan Polimis for assistance with data collection. We also thank participants in the PAA 2017 session on "Big Data Analysis in Demography" and participants in the IUSSP/ICWSM17 workshop on "Social Media and Demographic Research" for valuable feedback.

1 A summary of the advisory group's findings appears in the Documents section of *PDR* 41(1), March 2015.

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

3 This platform can be accessed without charge by anyone with a Facebook account at https://www.facebook.com/ads/manager/creation/creation/.

4 As of May 2017.

5 https://www.facebook.com/business/help/624074880953806

6 API, or application program interface: https://developers.facebook.com/docs/marketing-apis

7 We considered 52 supported countries and territories of origin 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, Venezuela, and Vietnam.

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

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

References

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

American Community Survey. 2014. United States Census Bureau. Available through IPUMS USA: https://usa.ipums.org/usa/

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

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

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

Ginsberg, J. et al. 2009. "Detecting influenza epidemics using search engine query data," *Nature* 457(7232): 1012–1014.

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

Herdağdelen, A. et al. 2016. "The social ties of immigrant communities in the United States," in *Proceedings of the 8th ACM Conference on Web Science*, pp. 78–84.

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

Pötzschke, S. and M. Braun. 2016. "Migrant sampling using Facebook advertisements: A case study of Polish migrants in four European countries," *Social Science Computer Review* 35(5): 633–653.

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

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

State, B., M. Rodriguez, D. Helbing, and E. Zagheni. 2014. "Migration of professionals to the U.S.—evidence from LinkedIn data," in *Social Informatics*, pp. 531–543.

State, B., I. Weber, and E. Zagheni. 2013. "Studying international mobility through IP geolocation," in *ACM International Conference on Web Search and Data Mining*, pp. 265–274.

Wang, W., D. Rothschild, S. Goel, and A. Gelman. 2015. "Forecasting elections with non-representative polls," *International Journal of Forecasting* 31(3): 980–991.

World Bank. 2015. International Migrant Stock. Data available via the Data Bank of the World Bank at: https://data.worldbank.org/indicator/SM.POP.TOTL

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

Wiśniowski, A., J.J. Forster, P.W. Smith, J. Bijak, and J. Raymer. 2016. "Integrated modelling of age and sex patterns of European migration" *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 179(4): 1007–1024.

Zagheni, E., V.R.K. Garimella, I. Weber, and B. State. 2014. "Inferring international and internal migration patterns from Twitter data," in *International World Wide Web Conference*, pp. 439–444.

Zagheni, E. and I. Weber. 2012. "You are where you e-mail: Using e-mail data to estimate international migration rates," in *Web Science*, pp. 348–351.

———. 2015. "Demographic research with non-representative internet data," *International Journal of Manpower* 36(1): 13–25.