7

Using Facebook and LinkedIn Data to Study International Mobility

CAROLINA COIMBRA VIEIRA, MASOOMALI FATEHKIA, KIRAN GARIMELLA, INGMAR WEBER AND EMILIO ZAGHENI

Introduction to advertising data

ONLINE SOCIAL NETWORKS are well known for promoting social interaction between people by connecting them even when they are physically separated, whether by a few miles or across continents. To deliver their free service to their users, most online social networks rely on targeted advertisements as their business model. Facebook, for example, hosts over a quarter of the world's population, and the Facebook Advertising Platform (Facebook Ads) alone is responsible for more than 98% of the company's total revenues. Similarly, Twitter, Instagram, and TikTok are social networks whose business models are based on advertising. Even social networks that are based on a freemium model such as LinkedIn, where many features are available for free, but some require a paid membership, rely on targeted advertisements via the LinkedIn Advertising Platform (LinkedIn Ads) for additional revenues.

The key appeal to advertisers who are considering using these platforms to promote their products lies in the targeting capabilities offered by the social networks. Highly targeted advertisements can, potentially, deliver the right message to the right consumer, and thus offer a good return on investment. Social networks are able to offer these targeting capabilities due to the rich user data they collect, which include detailed demographic information, such as information on the user's age, gender, home location, income level, and education level, but also on their topical interests and certain behaviours. Some of these attributes are explicitly self-declared by the users, such as their age and gender, while others are derived from meta-information on how they access the social network, such as their likely

Proceedings of the British Academy, **251**, 141–158, © The British Academy 2022.

home location or their device type; and some are inferred through machine learning models, using their likes, social interactions, or status updates as input.[1]

In other words, we can see social network advertising as the process of matching social network users by their profile data available on the social network to target groups specified by the advertiser to deliver advertisements. Thanks to the personalised targeting and scalability of online advertising, advertisers benefit from huge increases in their conversions and sales at a lower cost of acquisition. According to the non-profit Interactive Advertising Bureau, social network advertising has thus become a popular form of advertising, with a projected $43 billion spent on it in 2020.[2]

The richness of the aggregated data provided by online social network advertising platforms has also been explored by the academic community using the same tools the platforms provide to advertisers. These tools allow advertisers to obtain audience estimates referring to the estimated number of users on a social network that matches the given input criteria based on demographic attributes before the advertisement is launched. Using these tools, researchers can obtain data for studying demographic characteristics across several research areas, including mobility and migrant assimilation (Dubois *et al.*, 2018). Facebook and LinkedIn host two of the main advertising platforms researchers use to study migration, in part because these platforms have features that make it easy to target migrants.[3]

Facebook is the most popular online social network, and its purpose is very generic, as its users can utilise the platform for different purposes in their daily lives, such as contacting friends, sharing emotions, or reading the news. Similarly, users on LinkedIn can use the social network to connect and build a professional network. By using Facebook's advertising manager[4] it is possible, for example, to target users by their current location as well as by their home town, which allows us to identify potential migrants. Similarly, LinkedIn provides information on each user's previous and current place of work or study, which can be used to study migration. The information the network users provide about their native language and the languages they speak can also be used as a proxy for their nationality or home country, and to study migration.

When performing research involving Facebook Ads, it is possible to study, for example, the process of migration at a high level, and to examine the assimilation levels of migrants based on the interests they express online. For instance, in the 'Facebook examples' section we show how Facebook Ads can be used to collect anonymous and aggregate audience estimates for women aged 18+ living in Colombia who used to live in Venezuela and who primarily use a 4G connection to access

[1] Facebook: About detailed targeting, https://www.facebook.com/business/help/182371508761821?id=176276233019487

[2] IAB, https://www.iab.com/

[3] While Twitter has similar features, to the best of our knowledge there has so far been no research conducted using Twitter advertising data (at the time of writing, April 2021).

[4] Facebook Ads Manager, https://www.facebook.com/business/tools/ads-manager

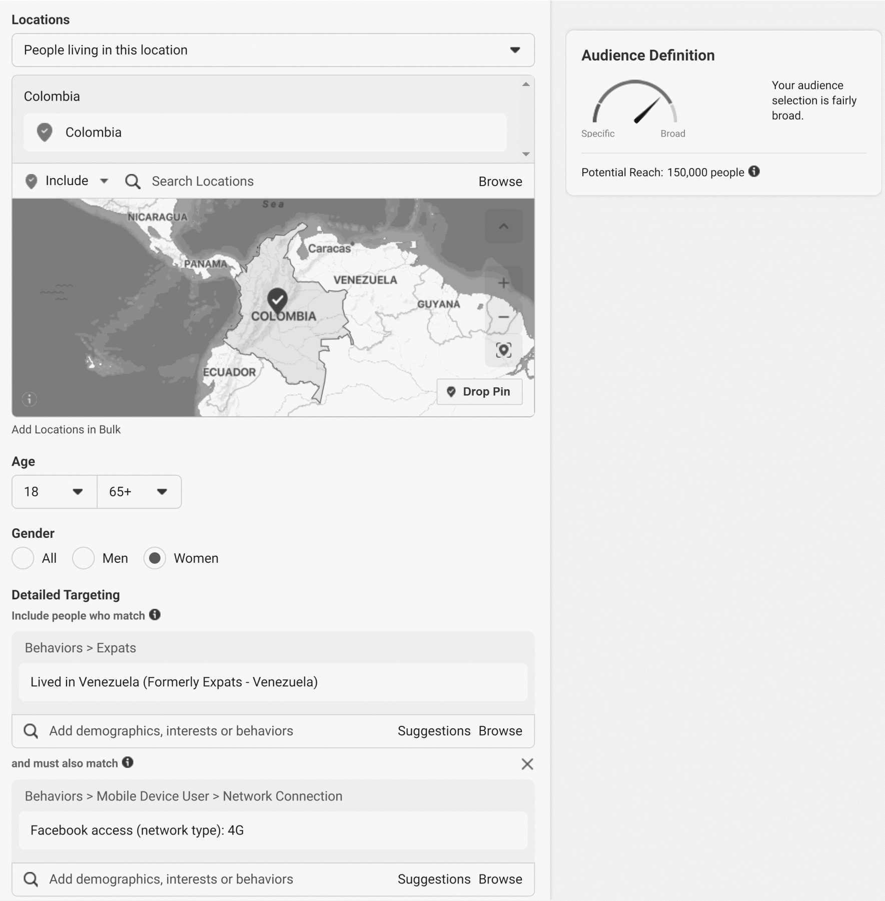
the social network site. Since it is possible to collect information about users' educational institutions or employers through LinkedIn data, LinkedIn Ads can be used to study professional migration. LinkedIn Ads may also be used to harvest information about where migrants have studied, where they live and work, and their professional connections. As an example, in 'LinkedIn examples' we show how LinkedIn Ads can be used to estimate the number of female LinkedIn members who hold a PhD, studied at the University of Cambridge, and are now living in Germany.

These data are publicly accessible (Facebook/LinkedIn account needed) for every user, and there are no special access requirements. The advertising platform web page allows for the most intuitive and user-friendly interactions. On this web page, the advertiser can select the desired demographic attributes, while the platform provides the audience estimates. However, this approach is relatively slow, and does not scale. To automate the data collection process we can use dedicated application programming interfaces (APIs) provided by the platforms. This approach requires certain levels of programming and web-scraping skills. These APIs, which are used for serving data to the website's front end, are not publicly advertised. They can, however, be easily identified using network monitoring in any modern web browser.

Thus, a key advantage of these data is that they are available for free. Another highly attractive feature of both platforms is that they allow researchers to segment the advertising and to target specific audiences (without actually placing an advertisement). When compared to the work involved in collecting data using other data sources, such as surveys and census data, the data collection process through online social network platforms requires less time, effort, and cost to deliver the desired data. The amount of data available, the scalability, and the speed of updates are among the other advantages of using these low-cost, real-time platforms. In particular, these platforms are interesting to researchers because they offer access to large numbers of users who provide their information on social networks, and because they offer APIs. Although there are a lot of advantages to using these platforms as data sources, doing so also raises ethical issues. In addition, the platforms are structured as a black box, which imposes limits on the use of such data for research. We discuss these limits in 'General limitations/challenges', after discussing examples from two major social networks that provide free advertising data: Facebook and LinkedIn.

Facebook examples

As is the case for most social networks, Facebook's main revenue stream is from targeted online advertising. Facebook offers advertisers a rich set of targeting options to reach the desired audience. In addition to providing advertisers with the ability to target users based on age, gender, device type, or topical interests, Facebook Ads allows advertisers to target users based on their previous country of residence ('users who used to live in country X'). For a specified set of targeting criteria, the platform then provides an estimate of the audience size called the Monthly Active



Locations

People living in this location

Colombia

Colombia

Include · Search Locations · Browse

Add Locations in Bulk

Age

18 · 65+

Gender

○ All ○ Men ● Women

Detailed Targeting

Include people who match

Behaviors > Expats

Lived in Venezuela (Formerly Expats - Venezuela)

Add demographics, interests or behaviors · Suggestions · Browse

and must also match

Behaviors > Mobile Device User > Network Connection

Facebook access (network type): 4G

Add demographics, interests or behaviors · Suggestions · Browse

Audience Definition

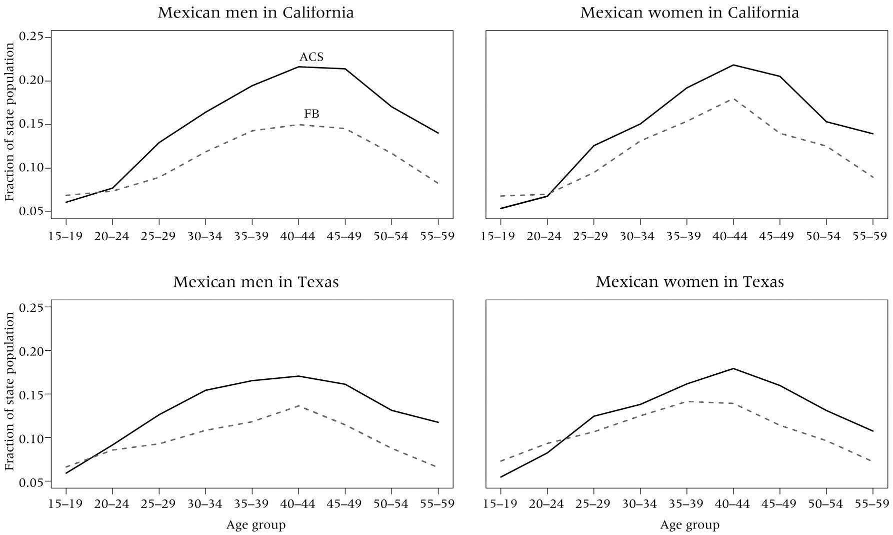
Your audience selection is fairly broad.

Potential Reach: 150,000 people

Figure 7.1 Screenshot of Facebook's advertising platform for an advertisement targeting women aged 18+ living in Colombia who used to live in Venezuela and who primarily use a 4G connection to access the social network site. The advertising platform displays an estimated audience of 150,000 users matching these targeting criteria.

Users (MAU). For example, Figure 7.1 shows the specifications for an advertisement targeting women aged 18+ living in Colombia who used to live in Venezuela and who primarily use a 4G connection to access the social network. For this target group, the advertising platform displays an estimated 150,000 users matching these criteria.[5] While the example here is at the country level, these estimates can be requested at various subnational spatial resolutions, including for regions, for

[5] Accessed on 10 November 2020. Note that the marketing API returns rounded estimates, whereby values in the thousands are rounded to the nearest hundred, values in the tens of thousands are rounded to the nearest thousand, and so on.



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

Figure 7.2 Facebook and American Community Survey (ACS) profiles of stocks of migrants by age and sex for Mexicans in California and in Texas. Source: Zagheni *et al.* (2017).

cities, and, at the lowest spatial granularity, for a specified radius around a given latitude/longitude coordinate. However, at more fine-grained spatial resolutions, data sparsity can become an issue, as the number of available Facebook users decreases, especially in countries with lower Facebook penetration. These estimates can be collected programmatically through an API[6] for various combinations of targeting options providing aggregated and anonymized data[7] on the distribution of Facebook users in a given location. Such data can provide opportunities for social science research, including for studies on migration. This section reviews some studies that have used these data for migration research, and the methods they have employed to analyse the data and to address bias.

Zagheni *et al.* (2017) and Spyratos *et al.* (2019) studied how migration researchers can complement migration statistics by using Facebook's audience estimates of users with different countries of previous residence to estimate the sizes of migrant populations. Zagheni *et al.* (2017) observed strong correlations between the fractions of Facebook users with various countries of previous residence across US states, the sizes of foreign-born populations estimated by the American Community Survey (ACS), and the country-level estimates of the sizes of migrant stocks in the World Bank data. When the authors compared the ACS data to Facebook's audience estimates, they observed patterns of bias across the information on age and country

[6] Meta for Developers: Marketing API, https://developers.facebook.com/docs/marketing-apis/

[7] The Python library https://github.com/maraujo/pySocialWatcher can be used to collect advertising data from the Facebook Marketing API.

of previous residence (Figure 7.2), which they corrected for by including country and age fixed effects in a regression model that predicted the survey data from the Facebook audience estimates. They observed improved model accuracy in estimating the sizes of migrant populations when this bias correction was done. Spyratos *et al.* (2019) used these data to compute estimates of migrant populations by country of origin globally in different destination countries. They addressed the selection biases in the sample of Facebook users by dividing the audience estimates for each migrant group by the Facebook penetration rate. This approach enabled them to correct for differences in Facebook usage across different countries of origin and destination, as well as across different age and gender groups. The Facebook penetration rates for each demographic group in each country were estimated by dividing the number of monthly active users on Facebook from the API by the respective population's size from the survey data. They then developed a model to compute the Facebook penetration level for each migrant group as a weighted average of the Facebook penetration rates in the countries of current and previous residence. These weight parameters were then estimated to give the best predictive fit between the corrected Facebook estimates and survey data.

Alexander *et al.* (2019) studied the utility of such data for estimating short-term population movements, such as in the aftermath of natural disasters, while focusing on the case of the outmigration of Puerto Ricans to the continental US after Hurricane Maria. They estimated the percentage change in the population of Puerto Ricans by comparing the number of Facebook users in the United States who previously lived in Puerto Rico in the periods before and after Hurricane Maria. In order to account for changes not due to the hurricane, such as changes related to an update to Facebook's estimation algorithm that may have affected audience estimates, they adopted a difference-in-differences approach whereby the observed change in the audience estimates for other migrant groups was deducted from the observed change in the audience estimates for Puerto Ricans. Given that changes in Facebook's audience estimates for other migrant groups generally moved in lock-step with those for Puerto Ricans in the periods prior to the hurricane, this approach was used to isolate changes in the population of Puerto Ricans that resulted from the hurricane. The authors then estimated the number of Puerto Ricans who moved to the continental US by multiplying the estimated percentage increase in the population of Puerto Rican Facebook users by the number of Puerto Ricans in the United States prior to the hurricane based on the survey data. The authors observed an increase in the number of Puerto Ricans living in the US, including a marked increase in the number of younger, working-age Puerto Ricans.

Palotti *et al.* (2020) studied how the estimates of Facebook users from Venezuela could be used to estimate the sizes of the Venezuelan refugee populations at the national and subnational levels for various countries in Latin America. To estimate the number of Venezuelan refugees, they corrected Facebook's audience estimates of users who previously lived in Venezuela by the overall Facebook penetration rate in the destination country. For example, if the total number of Facebook's monthly active users in the destination country, as per the advertising API, was estimated as equal to 60% of the destination country's population, then the raw estimates would be scaled up by a factor of 100/60. The authors found a strong correlation between these estimates and the available survey data at national and subnational geographic resolutions. In addition to estimating the sizes of the refugee populations, the authors explored the insights that could be gained from these estimates about the educational and economic situations of these populations. They developed regression models to estimate the users' income levels based on the device types they used on the social network, and drew on the self-declared information on the users' educational levels to learn more about the educational qualifications of the Venezuelan refugees across countries, and across regions in a country.

Beyond these quantitative estimates of the sizes of migrant populations, Dubois *et al.* (2018) and Stewart *et al.* (2019) studied qualitative aspects of cultural assimilation by using the interest-based targeting options of the platform. Dubois *et al.* (2018) collected the audience estimates of Facebook users for a list of interests in both the destination country and the home country of the target migrant group under study. They compared the numbers of users with these different interests in the home country and in the destination country in order to filter the interests to those that were more popular in the destination countries. Using the filtered list of interests, they then computed an assimilation score that compared the relative popularity of these interests in the destination country and among the targeted migrant groups in the destination country. Dubois *et al.* (2018) applied their approach to studying migrant assimilation among Arabic-speaking migrants in Germany across different demographic groups. Stewart *et al.* (2019) used the methodology of Dubois *et al.* (2018) to study the assimilation of Mexican migrants to the Anglo and African-American populations in the United States. For their study, Stewart *et al.* (2019) focused on the Mexican migrants' interests in musical genres in order to study their assimilation levels, while taking into account demographic dimensions such as age, gender, educational level, and language.

The wide range of targeting criteria provided on Facebook Ads makes it possible to study other aspects of migration and mobility as well. Spyratos *et al.* (2020) used data on estimates of users by travel frequency (using the option to target users who are 'frequent travelers' or 'frequent international travelers') to study the travelling behaviour of migrant groups by country of previous and current residence. As a validation step, the authors observed strong correlations between the estimates of Facebook users by country of previous residence who were frequent travellers or frequent international travellers, and the per capita number of international travellers by nationality in the United Kingdom, and the per capita income by nationality in the United States, respectively. To model the travelling behaviour of the migrants in this sample of Facebook users the authors fitted various regression models to explain the fraction of users who were frequent (international) travellers as a function of other variables, such as the users' demographics, and the characteristics of the country of previous or current residence, such as the country's income and gender inequality levels or its Facebook penetration rates.

As the studies reviewed here have shown,[8] the audience estimates from Facebook Ads represent a useful data source for studies of migration and mobility. Among the potential applications of these data are estimating the sizes of migrant populations; complementing traditional migrant statistics; estimating short-term migration movements in response to natural disasters, or in the context of a refugee crisis in which official data may be lacking or appearing with a time lag; and gaining further insights into other aspects of migration, such as education, socioeconomics, cultural assimilation, and travelling behaviour. However, these audience estimates also have limitations. 'General limitations/challenges' discusses the limitations of using social network advertising platforms for studies of migration.

LinkedIn examples

Unlike Facebook, LinkedIn's advertising platform does not provide a mechanism to explicitly target users based on countries they have lived in before. However, LinkedIn supports targeting users by 'Member Schools' to 'reach members who completed a course at a specific school, college, university or other learning institution'. This targeting is available for schools that have a dedicated page on LinkedIn, such as https://www.linkedin.com/school/university-of-cambridge/ for the University of Cambridge. Using the appropriate search API endpoint, one can look up the corresponding ID, in this case *urn:li:school:12691*. This ID is needed for API calls to obtain audience estimates through the corresponding endpoint.[9]

Figure 7.3 shows an example of the specifications for an advertisement targeting women living in Germany who studied at the University of Cambridge, and who hold a PhD. For this target group, the advertising platform displays an estimated audience of 530 users matching these criteria.[10]

To approximate the criterion of 'having lived in country X', we compiled a list of higher education institutions in country X, so that we could construct a query for LinkedIn users who 'studied at a higher education institution in country X'. Concretely, for each European country, including territories such as the Isle of Man, we (1) obtained a list of universities from uniRank[11] to search for, (2) searched for academic institutions for each city, region, or country targetable by LinkedIn, and (3) later filtered the returned results to remove kindergartens and high schools, as well as false positives, such as universities with ambiguous names from non-European countries (since we were only interested in European countries).

Figure 7.4 shows a circular plot of the estimate of LinkedIn users who (1) studied in country X (origin of the arrow), and (2) who now[12] live in country Y (target

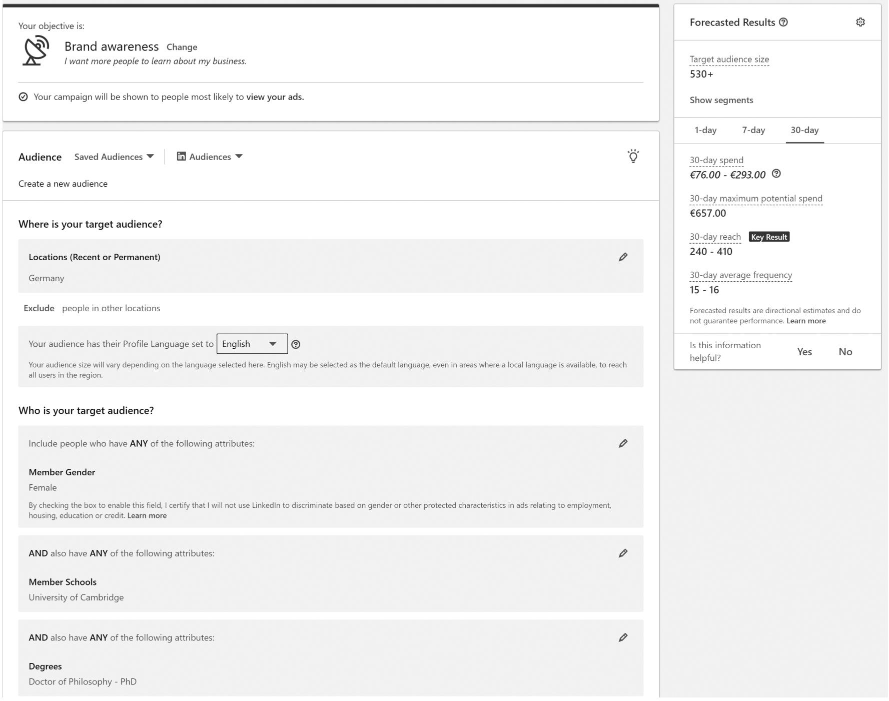
[8] Interested readers are invited to contact the study authors for access to the data used in their study.

[9] LinkedIn API Documentation, https://learn.microsoft.com/en-us/linkedin/

[10] Accessed on 20 April 2022.

[11] uniRank, https://www.4icu.org/

[12] As of the time of data collection, i.e., October 2018.



Your objective is:

Brand awareness Change

I want more people to learn about my business.

Your campaign will be shown to people most likely to **view your ads.**

Audience Saved Audiences ▾ Audiences ▾

Create a new audience

Where is your target audience?

Locations (Recent or Permanent)
Germany

Exclude people in other locations

Your audience has their Profile Language set to English ▾

Your audience size will vary depending on the language selected here. English may be selected as the default language, even in areas where a local language is available, to reach all users in the region.

Who is your target audience?

Include people who have **ANY** of the following attributes:

Member Gender
Female

By checking the box to enable this field, I certify that I will not use LinkedIn to discriminate based on gender or other protected characteristics in ads relating to employment, housing, education or credit. **Learn more**

AND also have **ANY** of the following attributes:

Member Schools
University of Cambridge

AND also have **ANY** of the following attributes:

Degrees
Doctor of Philosophy - PhD

Forecasted Results

Target audience size
530+

Show segments

1-day 7-day 30-day

30-day spend
€76.00 - €293.00

30-day maximum potential spend
€657.00

30-day reach Key Result
240 - 410

30-day average frequency
15 - 16

Forecasted results are directional estimates and do not guarantee performance. **Learn more**

Is this information helpful? Yes No

Figure 7.3 A screenshot of LinkedIn's advertising platform, showing a target audience size of 530 members for the selection of female LinkedIn users who studied at the University of Cambridge, hold a PhD, and are living in Germany.

of the arrow), rather than in country X. The audience estimates are in hundred thousands, and small countries are not shown to avoid clutter. Furthermore, the audience estimates are rescaled to correct for LinkedIn penetration in the target country: for a host country with a population of P and a total number of resident LinkedIn users L, all audience estimates are scaled upward by P/L. Assuming that migrants are more likely than non-migrants to be on LinkedIn, these rescaled audience estimates would be upper bounds for the true numbers.[13]

Whereas the fairly large country-to-country mobility shown in Figure 7.4 might not be surprising given the European context and programmess such as Erasmus,[14] the same methodology can be applied to shine a light on otherwise overlooked highly skilled migrants. Figure 7.5 shows similarly obtained estimates for LinkedIn users who (1) studied at a university in Syria, but (2) are currently[15] living in a

[13] Zagheni *et al*. (2017) observed that generally Facebook overestimates the (migrant)/(non-migrant) ratio compared to ground truth data from the World Bank. This overestimation was most pronounced in poorer, African countries. This suggests that migrants, compared to non-migrants, are more likely to be on Facebook and that this gap is biggest in poorer countries. In other words, if you migrate to a poor country, you are more likely to be on Facebook than the host population. But the same does not hold in countries such as the US or countries with high Facebook penetration.

[14] European Commission: Erasmus+, https://ec.europa.eu/programmes/erasmus-plus/node_en

[15] As of 2018.

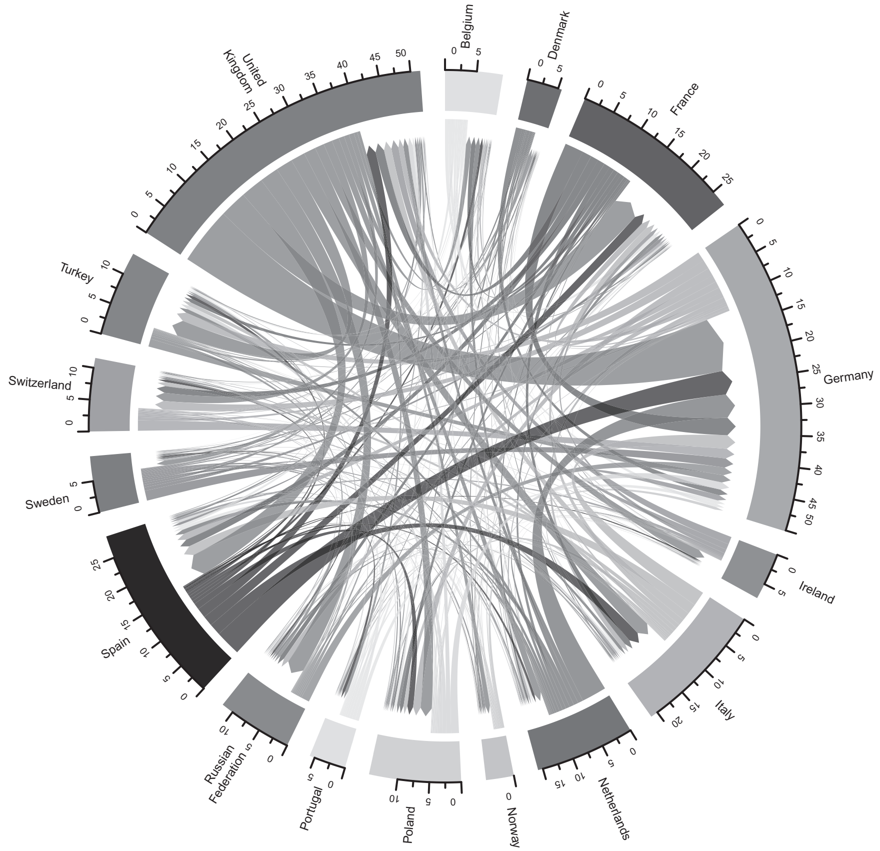


Figure 7.4 A circular plot showing estimates for the number of people who studied in one European country, but are currently living in another. The numbers are in hundred thousands (100,000s). The estimates were obtained through LinkedIn's advertising platform, and corrected for LinkedIn penetration rates in the host country. Assuming migrants are overrepresented on LinkedIn compared to non-migrants, the estimates would be upper bounds for the true numbers. Colour version printed as Plate 2.

European country. These audience estimates, which were generated by Michele Vespe and Spyridon Spyratos at the European Commission, were obtained by scaling up the raw audience estimates by the LinkedIn penetration among college graduates. Concretely, for each host country separately, they computed the ratio of 'LinkedIn users with a self-declared BA, MA, or PhD' and 'population with a university degree according to Eurostat[16]'. Assuming the LinkedIn penetration among university-educated Syrians is similar to the LinkedIn penetration among the host population, this correction factor appropriately corrects for the undersampling. This case study and the corresponding figure were shared by Vespe and Spyratos (2019), and are being used with permission.

[16] Eurostat, https://ec.europa.eu/eurostat

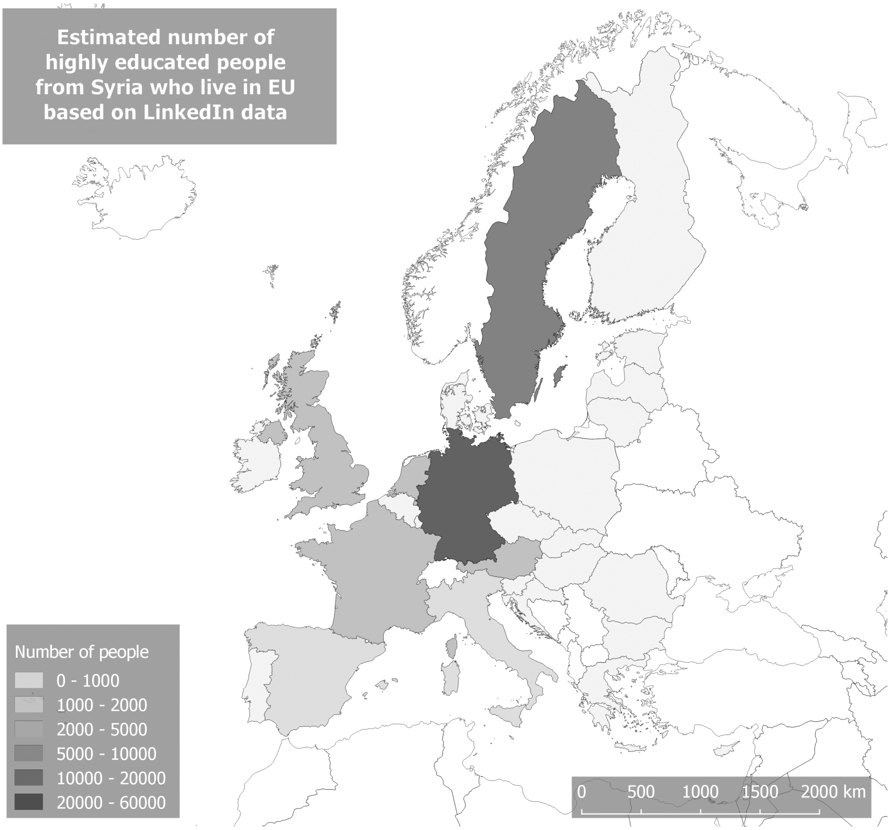


Figure 7.5 Estimates of the number of college-educated people from Syria living in the European Union (EU). The estimates were derived from LinkedIn Ads with an attempt to correct for selection bias (see text). A similar method was proposed by Vespe and Spyratos (2019).

Note that one challenge researchers face in this type of data collection is that LinkedIn users who studied in *several* countries are counted multiple times. Thus, a LinkedIn user who studied in both Germany and the United Kingdom, and who currently lives in Spain, would be counted both as having moved to Spain from the United Kingdom, and as having moved to Spain from Germany. This double count could conceptually be addressed by iterating over pairs or triples of origin countries. Concretely, a researcher could obtain a separate estimate for the number of people who have lived in two countries, such as Germany and the United Kingdom, and then subtract these estimates from the bigger count based on single countries of origin. While theoretically possible, this would increase the number of queries that would need to be issued to the appropriate API by roughly a factor of $n = 28$, which was the number of European Union member countries in 2018.

Another practical challenge researchers face in using these data is that, unlike Facebook, LinkedIn does not offer documentation for their existing free-of-charge APIs, which are somewhat hidden. The documentation offered online by LinkedIn instead refers to a premium API that is aimed at bigger advertisers. Therefore, we advise researchers who want to automate data collection from LinkedIn to study the hidden API using the network monitoring tools that are built into the browser.[17] These network monitoring tools log the requests sent to the LinkedIn server by the web browser. These requests are being made, usually unnoticed by the user, while the LinkedIn Campaign Manager is being used normally. Using these tools it is easy to understand which API call is being used to search for a school of a particular name, or which API call provides the actual audience estimates. Using the appropriate structure of API calls, together with the required cookies obtained after signing in, it is relatively easy to automate data collection using the APIs. In ‘General limitations/challenges’ we discuss the limitations and challenges that typically arise when working with advertising platforms.

General limitations/challenges

While there are many advantages to using these novel methods to obtain data for demographic research, it is important to be aware of their limitations and their risks.

Self-selection bias

The first limitation is integral to all social network research: namely, that the data are subject to self-selection bias. As most social network users self-select to be on the social network, they do not constitute a representative sample of the society. Certain features offered by the social networks also make them amenable to certain types of user sets. For instance, highly educated users are overrepresented on LinkedIn, and urban users are overrepresented on Twitter (Perrin and Anderson, 2019).

Self-reporting bias

Even if we correct for such biases, there are issues with using observational data that are subject to self-reporting bias. Users might not always provide the right information. For instance, in many cases we rely on self-reported data such as information about educational level on Facebook or about schools attended on LinkedIn. While it is generally assumed when using such datasets that the information users provide can be taken at face value, this may not always be the case.

Algorithmic bias

A more serious issue that is particular to advertising data is that these data were not devised to be used for the purposes we have illustrated in this chapter. First,

[17] See https://developer.mozilla.org/en-US/docs/Tools/Network_Monitor for Firefox, or https://developers.google.com/web/tools/chrome-devtools/network for Chrome.

these datasets were generated for advertising purposes, and not to study migration. Hence, the designers of these datasets might have different objectives. Second, the data are generated by black-box algorithms, which are, in most cases, company secrets that will not be revealed. Thus, any biases that were programmed into these black-box algorithms will be replicated in the data, and, eventually, in our analysis. Moreover, any errors in the inference processes of these algorithms will also be replicated in our analysis. Third, given the proprietary nature of the data and the processes through which they are generated, the data might undergo regular updates that are not apparent to the researchers, who mistakenly assume that the data collection process is consistent. For instance, Facebook has been changing its definition of 'users who lived in country X' every year. As a result, even though we see the ease of obtaining new data from these sources as positive (e.g., data for measuring how trends changed after an event, such as a natural disaster), we cannot be sure whether the changes in the data we observe are due to changes in Facebook's algorithms or to the new users who came onto the social network.

Brittleness of the APIs

From a practical perspective, the APIs researchers can use to obtain the data may change frequently or disappear altogether, which can make it difficult to replicate research. For instance, in our own experience, the LinkedIn advertising API changed from HTTP GET to HTTP POST unannounced, which forced us to rewrite all of our code.

Privacy concerns

The next big concern that arises when using such data is the privacy risks they pose. The datasets we proposed using in this chapter are aggregated anonymized data. They only provide counts of users who satisfy certain criteria. However, previous studies have demonstrated how such aggregated data and other tools that Facebook provides could be misused (Speicher *et al.*, 2018). Although these risks have now been fixed, there is no guarantee that the aggregated data will always be anonymized. To preserve individual anonymity, these advertising APIs do not return any user counts below a certain minimum threshold (1000 users on Facebook and 300 users on LinkedIn in November 2020). A practical side effect of the aggregation in research studies is that this aggregation could lead to biases. For example, accurate counts of small populations may not be possible. One technique for reducing these biases was recently proposed by Rama *et al.* (2020). It is even more important to bear in mind that having anonymized, aggregated data does not necessarily mitigate all privacy concerns. Privacy is not just of concern to individuals. Group-level harms can occur, even with aggregated data, that could be misused by bad actors. For instance, authoritarian governments could use such data to obtain aggregate statistics of regions in a country with high immigrant populations, or with people of a certain faith.

Ethical and legal concerns

Finally, there are ethical and legal challenges when using such data that remain unresolved. It is important to keep in mind that some of the data, particularly the data used to infer the interests of users, are collected by the platforms without the consent of the users. The use of data brokers (Venkatadri *et al.*, 2019) might lead to data being collected without the consent of the users, or without the users being given the opportunity to opt out. There are some interesting ongoing legal arguments about whether the computation and targeting of such interests complies with General Data Protection Regulation (GDPR).[18]

Summary

In summary, although advertising data might appear to be a viable source of data for demographic research, these data have specific caveats and limitations that we should be aware of when using such datasets in our research. This is especially important given that the output of such research could be used to inform policy. In these cases it is important to acknowledge biases, which may change the interpretation of results, but also to actively engage in bias mitigation. Box 7.1 summarises selected approaches for dealing with biases.

The next frontier: Combining data and fully leveraging the infrastructure of the digital age

In this last section of the chapter we discuss research developments in this area that may occur in the future. More specifically, we focus on the new opportunities offered by the ability to combine different types of data sources with those provided by the advertising platforms for Facebook and LinkedIn. The first main way of combining probabilistic samples and passively collected data involves developing statistical models, often using Bayesian approaches. The second main way of combining sources and tools involves using advertising platforms to recruit participants for surveys.

An emerging line of literature that we expect to develop further in the years to come focuses on combining different types of data sources using Bayesian statistical methods. The underlying idea is that different types of data have different types of

[18] Techcrunch: Facebook faces fresh criticism over ad targeting of sensitive interests, https://techcrunch.com/2018/05/16/facebook-faces-fresh-criticism-over-ad-targeting-of-sensitive-interests/

Box 7.1 Dealing with bias

There are different approaches that could be used to address different types of bias.

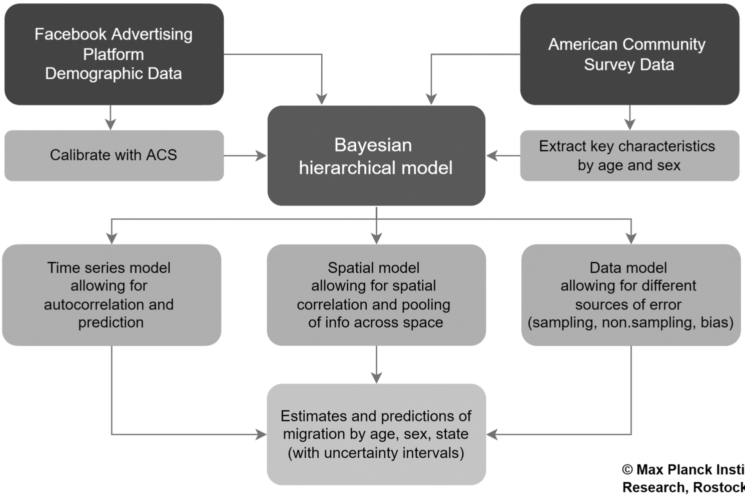
Settling for relative trends: Usually, relative trends, i.e., trends over time or differences across space, are somewhat robust against biases. For example, if globally only an unknown percentage $x\%$ of migrants are using Facebook, but if this percentage is stable across time and/or space, then an increase of 20% in the biased measure still corresponds to an increase of 20% in the true count. This general line of thought underlies the work of Palotti *et al*. (2020).

Machine learning and regression: Another approach is applicable if there is good ground truth to calibrate against and the aim is to extrapolate out to other locations or time points. Then the biased data can be used as features and the machine learning model takes care of the 'calibration' or 'de-biasing'. This approach assumes that the bias is a function that can be learned as any other function, as long as the machine learning model is suitable and there are enough training samples. This approach underlies the work of Spyratos *et al*. (2019).

Parametric approaches: If there are reasonable assumptions about how the bias behaves, it is possible to use parametric approaches. A structural form is derived from the related variables, and used to correct for the bias. For example, Zagheni and Weber (2012) analysed mobility patterns of email users in order to estimate population-level age and gender-specific migration rates, assuming a certain relationship between internet penetration and mobility to deal with selection biases. The profiles by age and gender were first rescaled to match official statistics, and then the estimated number of migrants, by age group and gender, was multiplied by a correction factor to adjust for overrepresentation of more educated and mobile people in groups for which the internet penetration was projected to be low.

Non-parametric difference-in-differences (DiD) approach: Without any type of data to calibrate against, this approach is a more formal version of trend analysis. But if there is some data to 'anchor' against then it becomes a kind of hybrid between machine learning and trend analysis. DiD involves comparing the difference in the size of a group of interest before and after an event to the difference in the size of a 'control' group before and after the same event. For example, Alexander *et al*. (2019) used it to model Puerto Rican migrant movements in the continental US before and after Hurricane Maria. Provided that the biases affect all groups equally, their effects are reduced.

Bayesian approaches: Bayesian formulation offers a coherent and probabilistic formalism to integrate various sources of uncertainty in the modelling, including the prior information and biases in the form of inflated or deflated values. The important parameters are modelled with statistical distributions, bias is adjusted over these parametrically, and prior distributions can be taken into account based on expert opinions or through (more) reliable data sources. Good examples are Raymer *et al*. (2013); Hsiao *et al*. (2020); Alexander *et al*. (2020) and, more recently, Rampazzo *et al*. (2021).


Facebook Advertising Platform Demographic Data
American Community Survey Data
Calibrate with ACS
Bayesian hierarchical model
Extract key characteristics by age and sex
Time series model allowing for autocorrelation and prediction
Spatial model allowing for spatial correlation and pooling of info across space
Data model allowing for different sources of error (sampling, non.sampling, bias)
Estimates and predictions of migration by age, sex, state (with uncertainty intervals)

© Max Planck Institute for Demographic Research, Rostock, Germany

Figure 7.6 Schematic illustration of the approach used to combine data from the Facebook advertising platform and data from the ACS to produce more accurate short-term forecasts of migrant stocks in the United States. Source: Alexander *et al.* (2020).

biases and imperfections. These issues, which can affect data quality, can be explicitly modelled by incorporating additional information. This can be done in a formal and reproducible way using Bayesian methods. More specifically, researchers can incorporate beliefs, either via the use of so-called priors, or by structuring models so that certain types of information are pooled across space and time.

In a recent example centred around nowcasting stocks of migrants in the United States, Alexander *et al.* (2020) showed that a Bayesian hierarchical model that combined data from Facebook Ads and data from the ACS produced more accurate short-term forecasts than either a model that relied only on Facebook data or a model that relied only on time series from the ACS.

Figure 7.6 shows a schematic illustration of the approach used to combine data from Facebook Ads and data from the ACS to produce more accurate short-term forecasts of migrant stocks in the United States. While the figure refers to the paper by Alexander *et al.* (2020), the approach is quite general, and a number of studies are testing Bayesian models for combining social network and survey data to estimate stocks or flows of migrants (e.g., Hsiao *et al.*, 2020; Rampazzo *et al.*, 2021).

The second main area of development is the use of the advertising infrastructure to collect data via surveys. Migrants are a hard-to-reach population. Targeted advertisements may provide the tools necessary to reach these populations and to invite them to participate in a survey. Initial approaches have demonstrated that this is a cost-effective way of collecting data (Pötzschke and Braun, 2017). A rapidly expanding related literature in public health has also shown that when appropriate post-stratification weights are used, surveys run via advertising platforms can generate estimates that are close approximations of those obtained from probabilistic samples (Grow *et al*., 2020).

In summary, we expect that the next frontier will involve blending approaches from classic data collections, survey methods, and Bayesian statistics with the infrastructure of the digital age, which offers not just access to digital trace data, but new opportunities for data collection and survey experiments.

Acknowledgements

We thank Carlos Callejo Peñalba, who was a master's degree student at Aalto University in Finland in 2018–2019. As part of his course project, Carlos collected and visualised LinkedIn data on the number of users who studied in one European country, but who now live in another European country.

We also thank Michele Vespe and Spyros Spyratos at the Joint Research Center of the European Commission in Ispra, Italy. Michele and Spyros let us use their analysis of the number of LinkedIn users who studied in Syria but who now live in a European country.

References

Alexander, M., Polimis, K. and Zagheni, E. (2019), 'The impact of Hurricane Maria on out-migration from Puerto Rico: Evidence from Facebook data', *Population and Development Review* 45(3), 617–30.

Alexander, M., Polimis, K., and Zagheni, E. (2020), 'Combining social media and survey data to nowcast migrant stocks in the United States', *Population Research and Policy Review* 41, 1–28.

Dubois, A., Zagheni, E., Garimella, K., and Weber, I. (2018), 'Studying migrant assimilation through Facebook interests', *in* S. Staab, O. Koltsova, and D. I. Ignatov, eds, *Social Informatics*, Vol. 2, Springer, New York, 51–60.

Grow, A., Perrotta, D., Del Fava, E., Cimentada, J., Rampazzo, F., Gil-Clavel, S., and Zagheni, E. (2020), 'Addressing public health emergencies via Facebook surveys: Advantages, challenges, and practical considerations', *Journal of Medical Internet Research* 22(12), e20653.

Hsiao, Y., *et al*. (2020), 'Modeling the bias of digital data: An approach to combining digital and survey data to estimate and predict migration trends', Technical report, Max Planck Institute for Demographic Research, Rostock, Germany.

Palotti, J., Adler, N., Morales-Guzman, A., Villaveces, J., Sekara, V., Herranz, M. G., Al-Asad, M., and Weber, I. (2020), 'Monitoring of the Venezuelan exodus through Facebook's advertising platform', *PLOS One* 15(2), e0229175.

Perrin, A. and Anderson, M. (2019), 'Share of US adults using social media, including Facebook, is mostly unchanged since 2018', Pew Research Center.

Pötzschke, S. and Braun, M. (2017), 'Migrant sampling using Facebook advertisements: A case study of Polish migrants in four European countries', *Social Science Computer Review* 35(5), 633–53.

Rama, D., Mejova, Y., Tizzoni, M., Kalimeri, K., and Weber, I. (2020), 'Facebook Ads as a demographic tool to measure the urban–rural divide', *in Proceedings of The Web Conference 2020*, 327–38.

Rampazzo, F., Bijak, J., Vitali, A., Weber, I., and Zagheni, E. (2021), 'A framework for estimating migrant stocks using digital traces and survey data: An application in the United Kingdom', *Demography* 58(6), 2193–218.

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), 801–19.

Speicher, T., Ali, M., Venkatadri, G., Ribeiro, F. N., Arvanitakis, G., Benevenuto, F., Gummadi, K. P., Loiseau, P., and Mislove, A. (2018), 'Potential for discrimination in online targeted advertising', *Proceedings of Machine Learning Research* 81, 5–19.

Spyratos, S., Vespe, M., Natale, F., Iacus, S. M., and Santamaria, C. (2020), 'Explaining the travelling behaviour of migrants using Facebook audience estimates', *PLOS One* 15(9), e0238947.

Spyratos, S., Vespe, M., Natale, F., Weber, I., Zagheni, E., and Rango, M. (2019), 'Quantifying international human mobility patterns using Facebook Network data', *PLOS One* 14(10), e0224134.

Stewart, I., Flores, R. D., Riffe, T., Weber, I., and Zagheni, E. (2019), 'Rock, rap, or reggaeton? Assessing Mexican immigrants' cultural assimilation using Facebook data', *in* L. Liu and R. White, eds, *WWW'19: The World Wide Web Conference*, Association for Computing Machinery, New York, 3258–64.

Venkatadri, G., Sapiezynski, P., Redmiles, E. M., Mislove, A., Goga, O., Mazurek, M., and Gummadi, K. P. (2019), 'Auditing offline data brokers via Facebook's advertising platform', *in* L. Liu and R. White, eds, *WWW'19: The World Wide Web Conference*, Association for Computing Machinery, New York, 1920–30.

Vespe, M. and Spyratos, S. (2019), 'A changing migration data landscape?', Global Working Group on Big Data for Official Statistics, International Meeting on Measuring Human Mobility.

Zagheni, E. and Weber, I. (2012), 'You are where you e-mail: Using e-mail data to estimate international migration rates', *in Proceedings of the Fourth Annual ACM Web Science Conference*, 348–51.

Zagheni, E., Weber, I., and Gummadi, K. (2017), 'Leveraging Facebook's advertising platform to monitor stocks of migrants', *Population and Development Review* 43(4), 721–34.