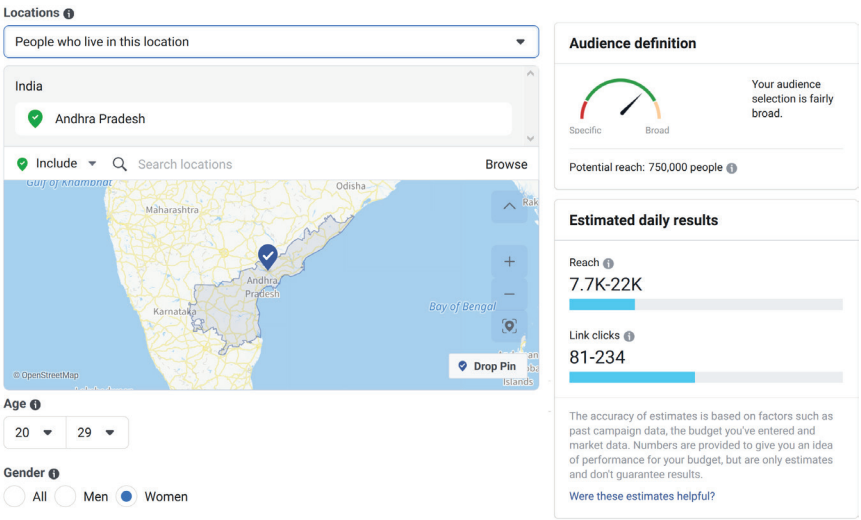
Using Facebook and Google Advertising Data to Measure the Gender Digital Divide

RIDHI KASHYAP (UNIVERSITY OF OXFORD), INGMAR WEBER (QATAR COMPUTING RESEARCH INSTITUTE), MASOOMALI FATEHKIA (QATAR COMPUTING RESEARCH INSTITUTE), IAN KNOWLES (UNIVERSITY OF OXFORD), AND REHAM AL TAMIME (UNIVERSITY OF SOUTHAMPTON)

Introduction

Sustainable Development Goal 5, achieving gender equality, highlights the importance of closing the global gender digital divide — that is, ensuring that women and girls have equal access to the Internet, mobile phones, and other digital technologies. However, the paucity and irregular production of data on these topics, particularly in less developed countries, makes monitoring progress towards this target difficult. In this brief, we show how anonymous, aggregated data from Facebook and Google's online advertising platforms can help fill the gap. We find that Facebook and Google advertising data are strongly correlated with gender gaps in internet access and digital skills. On our website digitalgendergaps.org, we take advantage of the better temporal resolution of the Facebook data to provide regularly updated indicators of internet and mobile use gender gaps globally, tracking digital gender inequalities as they evolve over time. Such a tracking system is much less expensive than conventional surveys and can help monitor the effect of policy interventions and economic shocks.

Figure 1. Facebook API showing the number of female Facebook users between the ages of 20-29 active in the state of Andhra Pradesh, India, in August 2019.



Our Approach

The Facebook and Google advertising platforms are designed to offer potential advertisers data on the group of users that they are interested in targeting. For example, the Facebook advertising platform can provide an answer to the question, "How many female Facebook users between the ages of 20-29 were active in Andhra Pradesh state of India in September 2019?" (Figure 1). When compared with data on men, this real-time information sheds light on the digital gender divide. In addition to age, gender, and location, we can also access information related to the device types, for example mobile devices, which are used to access Facebook. This is helpful for measuring different forms of digital access. Google's advertising platform, Google AdWords, which has an even broader global reach than Facebook, allows similar insights, though it differs slightly in showing estimates for ad impressions (the number of times an ad is seen by a user) rather than the number of users; more active users create more ad impressions. Similar information is provided by

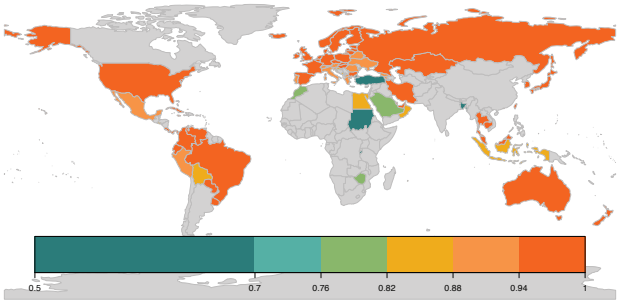
the advertising platforms of Twitter, LinkedIn, Snapchat, and others. The spatial resolution ranges from state level to sub-city postal codes, depending on country and advertising platform. To prevent reidentification of individual users, the platform limits audience estimates to no less than 100 users. This data can be used to construct gender indicators. We used the Facebook data to generate a "Facebook Gender Gap Index" (FB GGI), measuring the ratio of female to male monthly active Facebook users in a given country. For example, in Belgium we observed 3.6M female and 3.5M male monthly active Facebook users, whereas for India there were 65M female and 220M male monthly active Facebook users, as of September 2019. We examined how well this Facebook data, in combination with offline gender and economic indicators (e.g. the UN Human Development Index, gender gaps in education), predicted the latest available survey-based estimates of gender gaps in internet access and digital skills.

Results

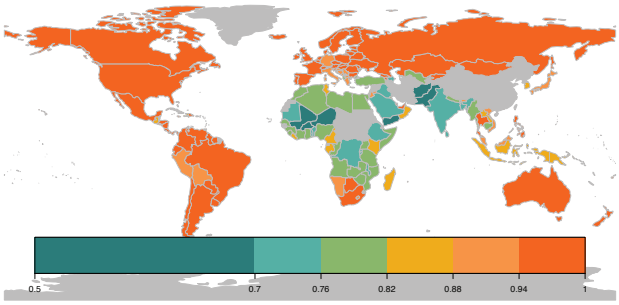
Our results show that both Facebook and Google online indicators are strongly correlated with survey data on internet access gender gaps and low-level digital skills (e.g., using copy and paste tools, transferring files, and sending email). When women are missing on these two online platforms, we can state with a high degree of certainty that they are not online at all, and lack digital skills in these countries. Although models that use Facebook indicators show better predictive performance than Google AdWords, models that combine Facebook and Google online data with offline development indicators perform best in predicting internet access gender gaps. In particular, the combination of Facebook and AdWords data and a country's Human Development Index explains about 80% of the variation in global internet gender gaps. Figure 2 shows how the global landscape of Internet access differs when viewed by survey data (2a) versus Facebook data (2b). Facebook indicators are better able to predict low-level digital skills compared with AdWords indicators. Our work highlights how women are disproportionately less online in countries in South Asia and sub-Saharan Africa compared with men, where values of the internet gender gap (the proportion of the female population with internet access divided by the male population with internet access) index lie between 0.7 and 0.8 (see Figure 2).

Figure 2. Two world maps showing the ratio of (percentage of women with Internet access)/(percentage of men with Internet access) on a per-country basis. ITU data from 2015 (top) is compared to model predictions of the online model using Facebook data from 2017 (bottom). The model manages to largely reproduce ITU ground truth data while substantially improving global coverage.

A) Internet access gender gaps according to 2015 ITU data.



B) Internet access gender gaps according to 2017 Facebook data.



Implications

Our research shows the great value of online advertising audience estimates in complementing existing traditional data sources on the lives of women and girls. All the data sources described here are publicly available free of cost, which enables near-real-time estimates to be made and, more generally, democratizes data access and analysis. When used ethically and responsibly, and in combination with existing data sources, online advertising audience estimates can help to fill gaps on important topics such as digital gender inequalities. These data streams could be used to provide crucial input into policies targeting populations with poor access to technology for infrastructural and educational investment.

Overall, however, this method has limitations. Perhaps most importantly, insights from online

data are more likely to represent the behavior of better-off people. However, this type of bias and data distortion is not necessarily problematic for two reasons. First, it is often exactly the missing data that is the signal. For example, in our research on gender gaps, the fact that women are not found in the data at the same rate as men provides a signal on gender inequalities. Second, approaches using supervised machine learning, such as regression models, treat the (biased) data merely as a signal to predict a particular quantity of interest, e.g. internet gender gaps derived from representative survey data. As long as the signal has high predictive power, it has potential value for the task. For such approaches, selection bias is only a challenge when it is non-systematic, e.g. when the reasons for bias differ across countries, and cannot be understood well enough to be corrected. On the other hand, if the reasons for bias are globally consistent, or if the contextual forces driving bias are well-understood, then estimates can be adjusted to account for the bias.

Another important challenge is the fact that the data provided to advertisers comes from proprietary methods. Academics and others cannot easily audit data quality. Whereas some user attributes such as age and gender are most likely derived from self-declared information, more detailed attributes — for example, Facebook's "lived in [country name]" category, which measures users who have previously lived in a given country but now live in another — are based on a proprietary inference algorithm with unknown accuracy.