Measuring global gender inequality indicators with large-scale online advertising data

Ridhi Kashyap[1], Florianne Verkroost[2], Reham Al Tamime[3], Masoomali Fatehkia[4], and Ingmar Weber[5]

[1,2](ridhi.kashyap,florianne.verkroost)@nuffield.ox.ac.uk, University of Oxford
[3]rat1g15@soton.ac.uk, University of Southampton
[4,5](iweber,mfatehkia)@hbku.edu.qa, Qatar Computing Research Institute

Keywords: gender inequality, digital trace data, big data for development, nowcasting, sustainable development goals

Extended Abstract

The promotion of gender equality features prominently in the United Nations Sustainable Development Goals (SDGs), both as a standalone goal (Goal 5) as well as in relation to other goals (e.g. access to education). Traditional data sources to measure progress on the SDGs are often outdated, lacking international comparability or appropriate disaggregation, or missing completely [4]. This paper demonstrates how anonymous, aggregate data from the online advertising platforms of LinkedIn and Google can be repurposed as a 'digital census' to measure and nowcast gender inequality indicators in skilled occupations and education globally. Although these data have widespread geographical coverage, are available in real-time or with high frequency, they are non-representative. We compute gender gap indicators using data from online populations on these platforms and compute their predictive performance by validating them against ground truth gender gap indicators computed from various data sources such as the International Labour Organization's Statistical Database, Global Gender Gap Report, and Wittgenstein Centre. We then explore the different types of biases of indicators generated using these online populations.

In the context of the United Nation's call for 'Data Revolution'[1] to close data gaps related to monitoring global development, this work contributes to understanding the complementary as well as shortcomings of an online data source for this endeavour [3]. While Facebook's advertisement platform has been used to study gender inequality in a global perspective [1, 2], to the best of our knowledge, this paper is the first to leverage LinkedIn and Google AdWords for a global analysis of gender inequality. This paper highlights how specific kinds of targeting unique to these platforms, such as targeting within skilled professionals in LinkedIn, or targeting based on filtered interests regarding education on AdWords can be used to capture specific domains of gender inequality.

Potential advertisers on online platforms such as Google or LinkedIn can specify a desired audience for their ads based on targeting criteria, such as gender, age, and geography. In this sense, they act as a type of census of the user base of these online populations. The estimates

[1] `http://www.undatarevolution.org`

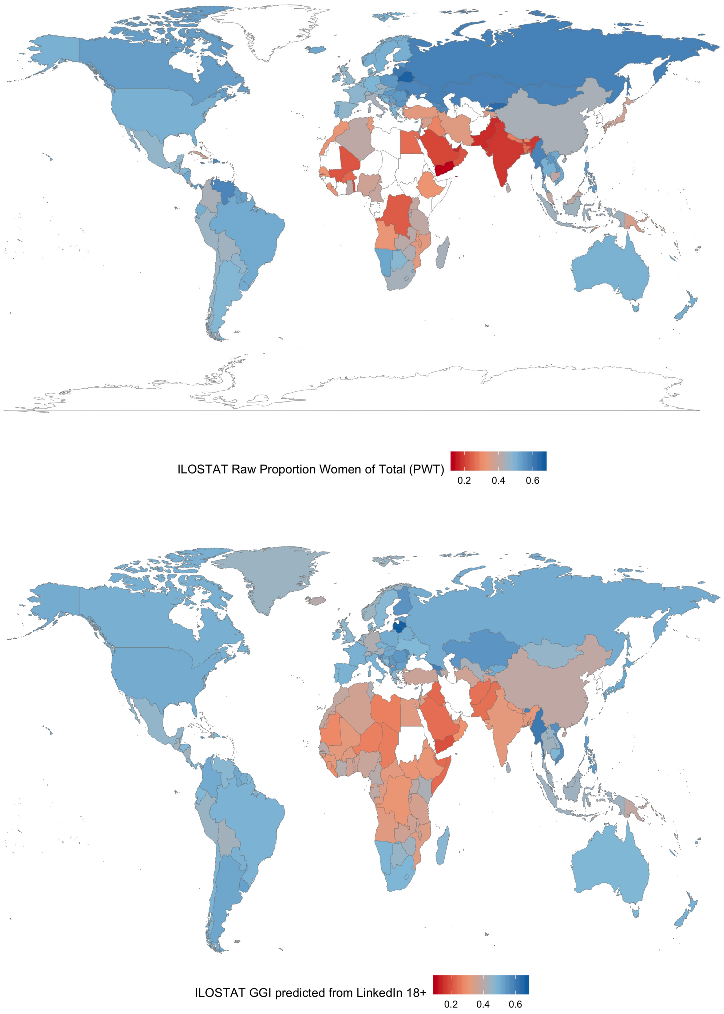
ILOSTAT Raw Proportion Women of Total (PWT)
0.2
0.4
0.6
ILOSTAT GGI predicted from LinkedIn 18+
0.2
0.4
0.6

Figure 1: (a, above) Gender gaps in skilled occupations computed using ILOSTAT (Skill levels 3 and 4), and (b, below) gender gaps in skilled occupations predicted using LinkedIn 18+ GGI.

available through the digital censuses of these online populations vary between platforms. For example, while LinkedIn provides *estimates of users* that match targeting criteria, the Google marketing platform, AdWords, provides *estimates of impressions*, which are the number of times an ad is expected to be shown. For LinkedIn, we collected data on LinkedIn users by gender, age and country, and computed female-to-male ratios of LinkedIn users for each country called the LinkedIn Gender Gap Index (LI GGI). We computed LI GGI indicators for each country across different dimensions of targeting, for example, by age, industry and seniority. Using AdWords, we collected data on the number of impressions disaggregated by age and gender for over 200 countries, further filtered for those who are actively searching and planning for post-secondary education. With these impressions we computed different AdWords Gender Gap Index (AdW GGI) variables for different age groups for each country, all of which were computed as female-to-male gender ratios of impressions.

In addition to these online GGI variables, our country-level dataset also included: 1) indicators of gender gaps in skilled occupations (ISCO levels 3 and 4, i.e. occupations requiring at least some post-secondary education and above) and share of female managers derived from the International Labour Organization's Statistical Database (ILOSTAT), 2) indicators of post-secondary educational enrolment, attainment and literacy gender gaps from the Global

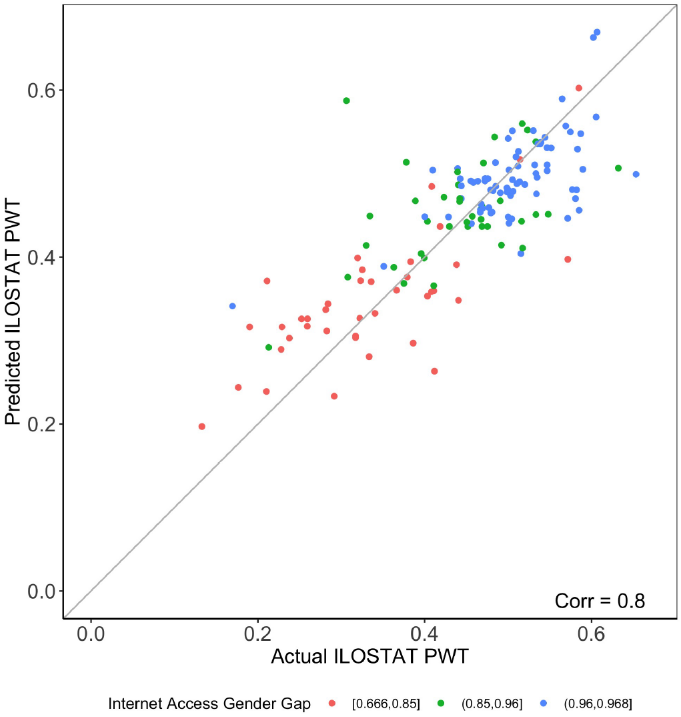

Figure 2: Ground truth occupational gender gaps from ILOSTAT (x-axis) and predicted gender gaps using LinkedIn GGI (y-axis). Points on scatterplot are color coded by internet access gender gaps.

Gender Gap report and the Wittgenstein Centre, 3) other offline indicators about a country's internet penetration and use, economic development, and demographic structure. We use the most highly correlated LI and ADW GGI indicators to predict gender gaps in skilled occupations from ILOSTAT and post-secondary education from the GGGR respectively using OLS regression models. We selected the most parsimonious one- or two-variable model using a step-wise forward selection procedure, and computed different measures of predictive fit (e.g. adj. R-squared, mean absolute error, SMAPE). To understand the biases of the indicators derived from these digital censuses we assessed which offline indicators best explain prediction errors.

Among different LI GGI indicators, we find that the LI GGI (for 18+ users), which is the female-to-male ratio of LinkedIn users, is most strongly, positively correlated with the gender gaps in skilled occupations derived from ILOSTAT (r = 0.79). The correlation of the LI GGI 18+ with the female share of managers in a given country is also positive, but weaker (r = 0.43). In contrast, correlations of AdW GGI indicators with ground truth educational gender gap indicators is weaker than those of LinkedIn gender gap indicators for skilled occupations. The correlations of AdW GGI for 18+ users with post-second education gender gaps are also positive, but weaker (r = 0.64 with post-secondary educational attainment gender gaps, r = 0.44 for post-secondary enrollment gender gaps).

The LinkedIn GGI 18+ provides the best performance, both in terms of adj. R-squared and out-of-sample performance in terms of SMAPE. A single-variable model with the LI GGI (18+) gives an adj. R-squared of 0.64 and SMAPE of 11.4% and improves geographical coverage of gender gaps in skilled occupations from 160 countries, for which data are available from the ILO, to 239 countries. Panel (b) of Figure 1 shows occupational gender gaps as predicted by LinkedIn compared with those computed using the ground truth data from the ILO. LinkedIn's predictions overcome data gaps in this indicator in Africa as well as South and West Asia. Interestingly, we find that LI GGI indicators are better able to explain variation in gender inequalities in skilled occupations in countries with greater gender inequality (low GGI values) (adj. R-squared = 0.43) than in high GGI (more equal) countries (adj. R-squared = 0.12). This suggests that when women are missing on LinkedIn, this is a strong signal that they are also missing in these countries in skilled occupations in the labour force, whereas in the more gender equal countries differentiated patterns of use of platforms may be apparent.

In contrast, single-variable AdW GGI models show weaker performance when predicting post-secondary education gender gaps compared with LinkedIn for occupational gender gaps (adj. R-squared 0.40, SMAPE = 22% for post-secondary attainment gender gaps, adj. R-squared 0.18 and SMAPE 23.08% for post-secondary enrollment gaps). This performance can be incrementally improved by combining AdW indicators across different age groups. In next steps, we will explore how the predictive fit of both LinkedIn and AdWords indicators can be improved by combining across multiple age groups and across platforms to predict gender gaps in both domains.

In preliminary analyses of the residuals of the predictive models using LinkedIn, we find that in countries with significant gender inequality in internet use (red points in Fig. 2), i.e where women are less online relative to men, our models using LinkedIn tend to under-predict gender inequality (above the x=y line in Fig. 2), or in others words, suggest that more women are in skilled occupations than that indicated by ground truth data. This indicates a selectivity in the female population that is online in these contexts, which is important for researchers to understand when using these data sources. In next steps, we will examine biases of predictions of gender gap indicators further and assess to what extent we can improve our predictions by correcting some of these biases.

References

[1] Fatehkia, M., Kashyap, R., and Weber, I. Using facebook ad data to track the global digital gender gap. *World Development 107* (2018), 189 – 209.

[2] Garcia, D., Mitike Kassa, Y., Cuevas, A., Cebrian, M., Moro, E., Rahwan, I., and Cuevas, R. Analyzing gender inequality through large-scale facebook advertising data. *Proceedings of the National Academy of Sciences* (2018).

[3] Letouze, E., and Jutting, J. Official Statistics, Big data and human development: towards a new conceptual and operational approach. Tech. rep., Data Pop Alliance and PARIS21, 2014.

[4] Weber, I., Kashyap, R., and Zagheni, E. Using advertising audience estimates to improve global development statistics. *ITU Journal: ICT Discoveries 1*, 2 (2018).