Predicting Digital Gender Gaps Using Facebook and Google Advertising Audience Estimates

Reham Al Tamime
University of Southampton
Southampton, UK
rat1g15@soton.ac.uk

Ridhi Kashyap
University of Oxford
Oxford, UK
ridhi.kashyap@nuffield.ox.ac.uk

Ingmar Weber
Qatar Computing Research Institute
Doha, Qatar
iweber@hbku.edu.qa

Masoomali Fatehkia
Qatar Computing Research Institute
Doha, Qatar
mfatehkia@hbku.edu.qa

Abstract

Gender equality in internet access and digital skills are important targets within the United Nations (UN) Sustainable Development Goals (SDGs). Measuring progress towards these targets is often challenging due to the limited availability of gender-disaggregated data on internet use and digital skills, particularly for less developed countries. In this paper, we examine how anonymous, aggregate data from the online advertising platforms of Google and Facebook can be leveraged to measure global digital gender inequality. Building on previous work that has used Facebook's marketing API, we assess the potential of another novel data source – Google's advertisement impression estimates (AdWords) – to measure digital gender gaps. AdWords provides estimates of the times an advertisement is shown on a search result page or another site on the Google Display Network. These estimates can be filtered based on targeting criteria such as age and gender. We generate gender gap indicators using both AdWords and Facebook data, and find that these online indicators are highly correlated with official statistics on gender gaps in internet use and low-level digital skills from the International Telecommunications Union (ITU) when available. We test different models using only online indicators, only offline development indicators, as well as those combining online and offline indicators to predict ITU digital gender gap measures. We find that the best performing models are those that combine Facebook and Google online indicators with a country's offline development indicators. Together with the HDI, the Facebook and AdWords gender gap indicators are able to explain about 80% of the variation in global internet use gender gaps. We highlight how appropriate regression models built on anonymous, aggregate, real-time data from online advertising platforms, can be used to monitor important global development indicators, with significant gains in geographical coverage for less developed countries. This is the first time that data from Google's Display Network, which claims to reach 90% of global internet users, is used for this purpose.

Introduction

The internet has revolutionized how individuals and communities seek information, communicate, and access goods and services. By lowering the cost of information and connectivity, the internet has tremendous potential to help meet sustainable development goals (SDGs) and this role is acknowledged in different SDG targets put forth by the United Nations (UN). Digital literacy forms an important part of the right to education (Goal 4).[1] The commitment to ensuring equitable access to the internet and other information and communication technologies (ICTs) is noted as a part of the goal for attaining gender equality (Goal 5), which pledges to "enhance the use of ... information and communication technology to promote the empowerment of women."[2]

Even as internet access has proliferated, 'digital divides' or inequalities in access and use of the internet persist (Scheerder et al., 2017; Robinson et al., 2015). Online inequalities often mirror socio-demographic, offline inequalities, and the digital divide by gender is one widely noted dimension of this inequality. According to the International Telecommunication Union (ITU), the UN's specialized agency for ICTs, over 250 million fewer women are online than men and gender gaps in internet use tend to be greater in developing countries (International Telecommunication Union, 2017).

The increasing visibility of the issue has led several UN agencies such as the ITU and UNESCO to endorse targets calling for gender equality in internet use and access to broadband (Broadband Commission, 2013; International Telecommunication Union, 2015; European Parliament, 2018). The lack of gender-disaggregated data on internet use however remains "one of the key barriers" in monitoring progress towards these development targets (Broadband Commission, 2013). Official, nationally representative gender-disaggregated statistics on internet use lack regular production and data availability on these indicators is especially limited in developing countries (Hafkin and Huyer, 2007). Data on gender gaps in specific ICT skills are available for even fewer countries than those for more general internet use measures. Routine survey data collection that collects information on individual level ICT use within households is expensive, and while some population censuses are able to collect information on internet or mobile availability at the household level, intra-household inequalities are not captured in these data sources (Fatehkia et al., 2018).

Given the challenges associated with regular data collection particularly in less developed country contexts, digital

[1] https://unstats.un.org/sdgs/metadata/

[2] https://sustainabledevelopment.un.org/content/documents/10789Chapter3_GSDR2016.pdf

trace data from the web have the potential to help fill this data gap and measure real-time gender disparities in internet use and ICT skills globally. Previous work has shown how Facebook's online advertisement audience estimates, which allow any user with a Facebook account to query the aggregate number of Facebook users by various demographic criteria, can be leveraged to predict gender gaps in internet use (Fatehkia et al., 2018).

This paper builds on the aforementioned study by examining the potential of another novel data source – Google's advertisement impression estimates (AdWords) – to generate real-time measures of these gender gaps globally. Whereas Facebook reaches a 'mere' 60% of Internet users[3], according to Google's own claims "the Google Display Network reaches 90% of Internet users worldwide"[4] Similarly to Facebook, Google allows advertisers to estimate the reach of their campaigns by showing them an estimate of the expected number of weekly impressions, i.e. the number of times an ad is expected to be shown on a search result page or another site on the Google Display Network. These estimates can be filtered based on different targeting criteria such as age and gender, and are available for over 200 countries.[5] Our paper examines the potential for AdWords, and compares the performance of AdWords and Facebook, both independently and together, for predicting gender gaps in internet use. Furthermore, we extend previous work by assessing what kinds of digital skill gender gaps the AdWords and Facebook indicators can help capture.

We generate a country-level dataset combining (i) online indicators on gender gaps derived from Google AdWords and Facebook's advertising audience estimates, (ii) the latest available statistics collected using surveys on gender gaps in internet use and different dimensions of ICT skills available from the ITU, and (iii) offline indicators related to a country's overall levels of development and gender gaps (e.g. education, occupations). With this dataset, we estimate models to predict ITU estimates of gender gaps in internet use and different ICT skills using both online indicators and a combination of online and offline indicators.

Our results show that both Facebook and Google online indicators are strongly correlated with ITU data on internet use gender gaps, as well as low-level ICT skills such as using copy and paste tools, transferring files, and sending emails. Although independently Facebook online indicators show better predictive performance than Google AdWords, the best performing predictive models are those combining Facebook and Google online indicators with a country's offline development indicators. We find that higher levels of human development, as measured by a country Human Development Index (HDI), are positively associated with greater gender equality in internet use. Together with the HDI, the Facebook and AdWords gender gap indicators are able to explain about 80% of the variation in global internet gender gaps. Our approach demonstrates how aggregate, anonymous advertising audience estimates from the two biggest online advertising platforms, can be leveraged to generate real-time measures of important sustainable development indicators linked to digital gender equality. This approach comes with big gains in geographical coverage for less developed countries where existing data are often lacking or infrequently collected. As much less is known about patterns of use, and in particular gender gaps in use, of social media and other online platforms in less developed country contexts, our work contributes towards developing and understanding new measures of global digital gender inequality.

[3] 2.2B out of 3.6B – see https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/ and https://www.statista.com/statistics/273018/number-of-internet-users-worldwide/.

[4] https://ads.google.com/home/how-it-works/display-ads/

[5] https://support.google.com/google-ads/answer/6320

Background

Digital Gender Gaps

Socio-demographic, economic, cultural and motivational factors have been shown to affect gender gaps in internet access and use (Scheerder et al., 2017). Evidence gathered from developed countries has shown that the gap between men and women has closed in terms of internet access as internet penetration has increased. However, a 'second-order' digital divide in terms of patterns of use and skills has been found with women showing lower frequency of use, a narrower range of online activities and lower likelihood of reporting strong internet skills, even if their actual web use skills were not lower than those of men (Robinson et al., 2015; Ono and Zavodny, 2007; Hargittai and Shafer, 2006; DiMaggio et al., 2004). These findings have led scholars to suggest that the the digital divide is better conceptualized in terms of a spectrum of skills instead of a binary classification of whether an individual is an internet user or not (Hargittai, 2002). Digital skills are important to understanding whether the benefits of the internet are accrued evenly or whether digital inequality is further exacerbated as internet penetration increases (Hargittai and Shafer, 2006).

In developing countries, the percentage of women using the internet lags behind the percentage of men using the internet across all age groups (Antonio and Tuffley, 2014). Internet access gender gaps in developing countries reflect broader structural inequalities in terms of access to education, employment and income that women face (WWW Foundation, 2015; Robinson et al., 2015; Hilbert, 2011; Hafkin and Huyer, 2007). In addition to these socio-economic barriers to access, studies have also documented how cultural norms in patriarchal contexts may also impede women's internet use, particularly when internet access is mediated via men (Abu-Shanab and Al-Jamal, 2015; WWW Foundation, 2015; Gurumurthy and Chami, 2014; Intel, 2012). The discussion of the digital divide in the context of developing countries has focused largely on access or general use inequalities, as survey data on digital skills or digital literacy in these country contexts are especially limited.

Monitoring Development Indicators with 'Big Data'

In recent years, several researchers and international organizations have begun to explore the potential for 'big data' sources to overcome challenges associated with limited data coverage on development indicators, particularly in developing countries (IUSSP, 2015; IEAG, 2014; Letouze and Jutting, 2014). These works have used diverse big data sources, with examples ranging from the use of mobile call log data to predict income in African countries (Blumenstock et al., 2015; Mao et al., 2015), to night-time satellite data to measure poverty (Elvidge et al., 2009), to web search and public social media posts to predict unemployment and health outcomes (Resce and Maynard, 2018; Nuti et al., 2014; Choi and Varian, 2012). Despite weaknesses in big data sources, such as issues of non-representativeness and limited metadata to understand the data-generating process, a significant strength of these data sources is their (near) real-time measurement, which make them promising for 'nowcasting' (Salganik, 2017; di Bella et al., 2016). Nowcasting is typically employed when the actual value of indicator of interest will only be known with a significant delay, creating the need to "predict the present" (Choi and Varian, 2012).

One of the big data sources used in this study, Facebook's advertisement audience estimates available from the platform's marketing API, can be queried for information on the number of Facebook users by various demographic characteristics and can be thought as a kind of real-time census over the platform's user base. These data have been leveraged to study population health (Araújo et al., 2017; Chunara et al., 2013), to provide demographic estimates of migration (Zagheni et al., 2017) and male fertility (Rampazzo et al., 2018), and to generate gender inequality measures, including most relevantly for this study country-level internet gender gaps (Fatehkia et al., 2018). Gender gaps in Facebook use across countries have also been shown to be correlated with different domains of gender inequality more generally, including education, health and economic opportunity (Garcia et al., 2018). In countries where gender inequalities in socio-economic domains is larger, men also outnumber women on Facebook (Fatehkia et al., 2018; Garcia et al., 2018). In large countries such as India, significant sub-national variation in gender gaps in Facebook use exist and these data have been used to generate sub-national measures of digital gender inequality. Some of the variation in sub-national digital gender inequality can be explained by differences in socio-economic development between Indian states, with states with higher GDP per capita, literacy and internet penetration showing less skewed gender gaps in Facebook use (Mejova et al., 2018).

Our work applies the approach developed in Fatehkia et al. (2018), who developed an indicator called the 'Facebook Gender Gap Index' and found it to be highly correlated with ITU statistics on internet gender gaps collected using surveys fielded by national statistical agencies. The authors further found that predictive performance of the Facebook indicator was enhanced when combined with offline indicators linked to a country's development and offline gender gaps. We expand their work by (i) going beyond internet use gender gaps to also explore correlations with ICT or digital skills gender gaps, and by (ii) including data from Google, the biggest online advertising platform, to build more accurate models.

To the best of our knowledge, ad impression estimates for Google's Adwords platform have not yet been used to monitor and model online gender gaps or any other targets related to the SDGs.

Data

Our dataset comprises (i) online indicators derived from advertisement impression estimates available from Google AdWords and advertisement audience estimates from Facebook's Marketing API, (ii) an extensive range of offline indicators related to a country's level of development (e.g. GDP per capita, Human Development Index), and gender inequalities (e.g gender gaps in literacy), and (iii) ITU data on internet use as well as specific ICT or digital skills by gender and country of the user collected using nationally-representative surveys, which we use to derive our ground truth internet use or digital skills gender gap indicators. These different indicators are described in this section.

Online indicators: AdWords Gender Gap Index and Facebook Gender Gap Index

The online data we use come from publicly-accessible, anonymous and aggregate data that are provided to advertisers by online platforms to estimate the potential reach or audience size of their advertisement campaigns. The Facebook's advertisement audience estimates used in Fatehkia et al. (2018) are the number of monthly active Facebook users (MAUs) disaggregated by various geographic and demographic attributes, such as user age and gender. In contrast, the AdWords' data do not provide information on the numbers of users but instead the number of *impressions*. Impressions are counted each time an ad is shown on a search result page or other site on the Google Network. The provided impression counts are weekly estimates and can be disaggregated by various geographic and demographic attributes. As Google allows advertisers to create different kinds of campaigns on AdWords including search network only, display network only, video, shopping and universal app, we selected the display network campaign in order to retrieve the number of impressions for this study.

AdWords' reach estimates have the potential to be used for social science research questions pertaining to the attributes of the world's population, as they capture information related to the world's online population in real-time. Even though the use of AdWords impressions for social science applications has been limited, Facebook's advertisement audience estimates have been used in a similar vein as a type of 'digital census' (Zagheni et al., 2017; Fatehkia et al., 2018). Zagheni et al. (2017) found that the Facebook data are able to provide good demographic estimates of quantities such as percentages of the population of a particular nationality. As Google is more widely visited than Facebook[6] our

[6] `https://www.alexa.com/topsites` and see foot-

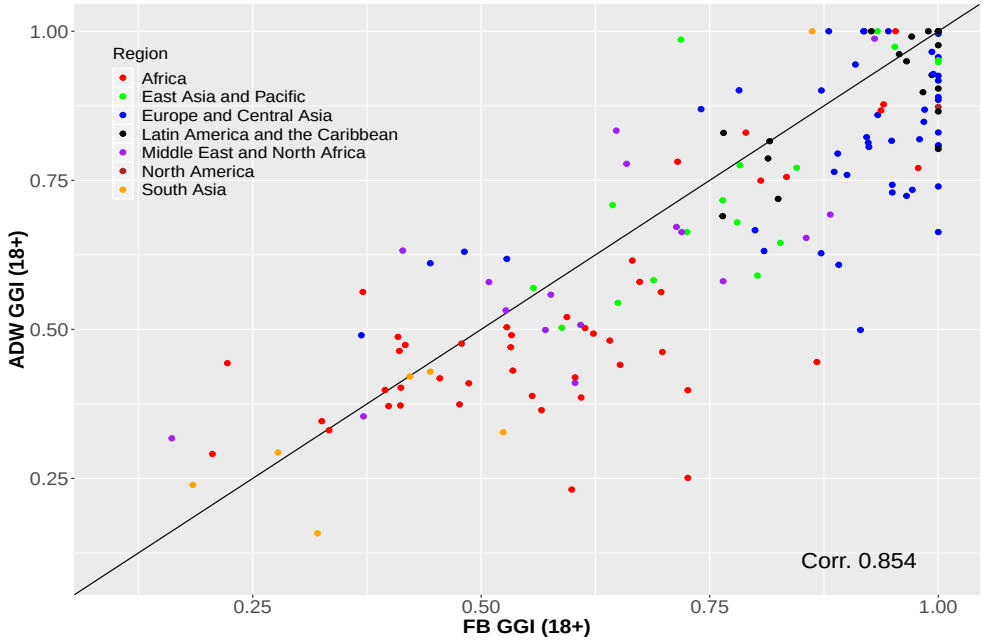

Figure 1: The FB GGI (ages 18+) against the ADW GGI (ages 18+). Each point indicates a country and points are colour coded by world region. The line is the x=y diagonal.

study explores the use of AdWords' number of impressions as a novel data source to monitor digital gender gaps. An important difference from Facebook's user estimates to note is that more active users are likely to cause more impressions, potentially leading to certain biases. For the purpose of this study, the number of impressions are disaggregated by age and gender and collected in June 2018 from 200 countries.[7] Also, we selected the language to include all languages spoken in every country. The 'unknown' number of impressions were excluded from the analysis as they related to audience whose age and gender have not been identified. The number of impressions were used to compute the AdWords Gender Gap Index (ADW GGI) for each country as follows:

$$\text{ADW GGI} = \frac{\text{Female to male gender ratio of impressions}}{\text{Female to Male gender ratio of the population}} \quad (1)$$

We divide the gender ratio (female to male) of impressions on AdWords with the population gender ratio of the same age category as the AdWords ratio. This is to correct the AdWords-derived measures for population imbalances. For example, countries such as the United Arab Emirates (UAE) have a much larger male than female population due to influx of foreign male workers. Correspondingly, observing a gender imbalance in terms of number of Google impressions (or users) for the UAE could be merely reflecting the (offline) population gender imbalance. We obtained the population gender ratios for various age groups from the UN World Population Prospects Database (United Nations Population Division, 2017).

Similarly, The Facebook's advertisement audience estimates were collected in May 2018 for 193 countries from Facebook's Marketing API.[8] The data were on monthly active Facebook users disaggregated by age, gender and country. The Facebook Gender Gap Index (FB GGI) for each country can be defined as:

$$\text{FB GGI} = \frac{\text{Female to male gender ratio of users on Facebook}}{\text{Female to Male gender ratio of the population}} \quad (2)$$

We have excluded some countries from the analysis as Google AdWords provides vague estimates for those countries, such as "$>$ 1B", which does not permit us to calculate ratios. Countries affected by these vague estimates include Aruba, Brazil, Czech Republic, Hong Kong, India, Japan, Macedonia, Micronesia, Turkey, and the United States. The Facebook data are not affected by this issue. Furthermore, some countries were excluded because they are listed under AdWords as countries such as "Saint Pierre and Miquelon", but not recognized by the World Bank or the UN. From the Facebook data, we excluded countries with less than 1 million users. After imposing these restrictions, we were left with data for 176 countries from Facebook and 166 countries from AdWords for the age group 18+. Both Google and Facebook support further filtering impressions/users with certain characteristics. In particular, we compute variants of

note 4]

[7] https://developers.google.com/adwords/api/docs/appendix/geotargeting

[8] Information and documentation about Facebook's Marketing API is available here: https://developers.facebook.com/docs/marketing-apis.

Table 1: Data availability, features, and correlations of the different variables with the ITU Internet Gender Gap Index

Variable	Number of Countries in Dataset	Pearson's Correlation with ITU Internet GGI	Year
ITU Internet GGI	83	1	Varies (2013-17)
Internet Penetration	188	0.695	Annual (2016-17)
log(GDP per Capita)	173	0.680	Annual (2016)
HDI	179	0.737	Annual (2015)
Mean Years of Schooling (HDI)	179	0.630	Annual (2015)
FB GG age 18+	176	0.731	Real time (2018)
FB GG age 20-64	176	0.734	Real time (2018)
FB GG age 25+	176	0.721	Real time (2018)
FB GG age 25-29	177	0.726	Real time (2018)
ADW GG age 18+	172	0.622	Real time (2018)
ADW GG age 25+	172	0.608	Real time (2018)

Table 2: Correlations of the different ICT Skills Gender Gap Index with Facebook and AdWords Gender Gap Index

	Number of Countries in Dataset	FB GG age 18+	Number of Countries in Dataset	ADW GG age 18+
DS GGI-Copying or moving file or folder	46	0.731	44	0.604
DS GGI-Using copy and paste tools	32	0.803	31	0.700
DS GGI-Sending e_mails with attached files	16	0.841	15	0.761
DS GGI-Using basic arithmetic formula in spreadsheet	41	0.550	39	0.479
DS GGI-Connecting installing new devices	19	0.179	17	0.377
DS GGI-Finding, downloading and installing software	39	0.452	37	0.393
DS GGI-Creating electronic presentations	46	0.763	43	0.665
DS GGI-Transferring files between computer and a device	47	0.767	44	0.602
DS GGI-Writing computer program using programming language	46	-0.248	43	0.013

ADW GGI and FB GGI for different age groups of users (e.g. 25-29 year olds).

All gender gap indicators used in this study take values greater than zero, with values less than 1.0 indicating countries where there is a gender gap with women at a disadvantage. Values close to 1.0 are desirable as they indicate gender parity. This is in line with the methodology of the Global Gender Gap Report (GGGR) published by the World Economic Forum (World Economic Forum, 2016). In this sense, a higher value on this index indicates a gender gap that has been closed. Also in line with the GGGR, we truncated values larger than 1.0, corresponding to women doing better than men, as our focus is on gender equality rather than women's empowerment. Hence our measure does not differentiate between countries that have attained parity (a gender gap index equal to 1.0) and those where women have surpassed men (a gender gap index greater than 1.0).

Fig. 1 shows the values of the FB GGI (18+) and the ADW GGI (18+) for all countries for which both indicators are available colour coded by the region of the world they are in. Countries in Africa and South Asia show the largest gender gaps, followed by those in North Africa and the Middle East. Overall, as also shown in the same Figure, the FB GGI and ADW GGI for countries are strongly correlated with each other, with a Pearson's correlation coefficient of 0.85. On average, there are a greater number of points below the x=y diagonal than above it. This suggests the same countries tend to score higher on the FB GGI, indicative of greater gender equality, than the ADW GGI. Although this could potentially indicate differential patterns of use of Google and Facebook by gender, this could also reflect the fact that the ADW GGI is based on impressions whereas the FB GGI is based on users. If men are more active users of the internet and generate a greater number of impressions or page views on Google, then this is likely to result in greater gender gaps in the ADW measure relative to the FB measure.

Dependent Variable: Internet Use Gender Gap Index

Although both Google and Facebook have large numbers of users, not all internet users use these platforms. Thus, to evaluate to what extent a gender gap index computed using the audiences of these platforms captures internet use gender gaps as well as gender gaps in other digital skills, we need to correct our online indicators of potential biases. To do this, we need to compare and validate the online indicators against indicators against ground truth measures of internet use by gender and location. We use data on internet use by gender of the user and country reported in the ITU World Telecommunications Indicators Database (International Telecommunication Union, 2018).

The ITU data provide proportions of individuals using the internet by gender to give gender-specific internet penetration rates (e.g. 40% of women in a given country are internet

users). The data are collected using nationally representative surveys fielded by national statistical agencies in the ITU's member states. The latest edition of the data covers surveys that were fielded in member states between 2013 and 2017. The data are thus available for different years for different countries based on whenever the survey was fielded there. The Internet Use Gender Gap Index (Internet GGI) for a country using ITU data is defined as:

$$\text{Internet GGI} = \frac{\%\text{ of female population using internet}}{\%\text{ of male population using internet}} \quad (3)$$

In addition to data on internet use by gender, the ITU also provides data on different types of ICT skills by gender for different countries (International Telecommunication Union, 2018). These measures capture a range of specific skills such as being able to copy or move a file or folder, sending e-mails with attached files, using basic arithmetic formula in a spreadsheet, among others listed in Table 2. These measures are collected via a survey, and are self-reported in response to questions that ask whether the user has undertaken the specific computer-related activity in the last three months (International Telecommunication Union, 2016). Some but not all of these indicators capture specific skills related to internet use. The number of countries for which the skills data are available are considerably fewer than those for which internet use by gender is available (see Tables 1 and 2). Where data are available, we can also compute a digital skill-specific GGI (DS GGI) akin to the Internet GGI. However, due to sparsity of ground truth for training purposes, our ability to use all of these measures as dependent variables in predictive models is limited.

Offline Predictors

Our dataset includes offline predictors that have been associated with internet use gender gaps (Fatehkia et al., 2018). These include factors associated with a country's overall levels of development captured in indicators such as the GDP per capita and different dimensions of the Human Development Index (HDI), its ICT infrastructure captured in indicators such as the internet penetration rate, and gender-specific development indicators such as, for example, gender gaps in literacy. In addition to these economic and development indicators used in previous work (Fatehkia et al., 2018), we also include six cultural variables drawing on Hostede's cultural dimensions theory, which classify different countries based on six dimensions: power distance, uncertainty avoidance, individualism/collectivism, masculinity/femininity, long/short-term orientation and indulgence/restraint (Hofstede, 2011).

Methods

We fitted three types of ordinary least squares (OLS) regression models to predict the internet GGI, our ground truth measure of internet use gender gaps. These include: (i) online models using different ADW GGI and FB GGI predictor variables, (ii) online-offline models to assess if some of the biases in using only online indicators for measuring internet use gender gaps could be corrected when combined with offline variables, and (iii) an offline model in which only offline predictors were used. We also attempted to use FB and ADW GGI variables from different age groups in our online models. For models that relied on multiple variables, we performed variable selection using a greedy stepwise forward approach, whereby variables were iteratively added to the model, starting out with a model with just an intercept, in order to increase the adjusted R-squared of the resulting model (Fatehkia et al., 2018). The offline models are a benchmark against which we compare the predictions of the online models, which rely on data from Google and Facebook to generate insights about internet use gender gaps. Due to the smaller number of countries for which ITU ground truth data are available for the digital skills GGI, we estimated the best-fit two variable models using either online or offline indicators when predicting these outcomes. To evaluate the performance of different models for predicting the internet and DS GGI, we report three measures of model fit: (i) Adjusted R-squared, (ii) Mean Absolute Error, and (iii) Symmetric Mean Absolute Percentage Error (SMAPE). For the first one, larger values and values closer to 1.0 indicate better performance. For the last two, lower values and those closer to 0.0 indicate better performance. The SMAPE is computed using a Leave-One-Out cross validation procedure where the model is fitted on all of the data except for one country, and the fitted model is then used to predict the left out data point.

Results

Correlation Analysis

Table 1 presents different indicators that are the most strongly correlated with the internet GGI. Among offline indicators, the internet penetration rate, log GDP per capita, HDI and mean years of schooling are positively correlated with the internet GGI.

The AdWords and the Facebook GGI variables for different age groups are also strongly correlated with the internet GGI. The ADW GGI for ages 18+ has a correlation value of 0.622 and for ages 25+ a value of 0.608 with the internet GGI. The FB GGI measures show a stronger correlation with the internet GGI measures than the ADW GGI ones, with the FB GG for age 18+ and ages 20-64 showing the strongest correlations.

Table 2 presents correlations of the FB and ADW GGI with digital skills gender gap measures from the ITU (DS GGI). The strongest skill-specific measure that is correlated with the FB and ADW indicators is linked to an internet-specific skill of sending emails with attached files, although this is available for only 16 countries, followed by skills linked to using copy-paste tools. These results suggest that Facebook and AdWords gender gap indices could also to capture low-level digital skills. For high level skills such as programming, the correlation with FB and ADW GGI measures is much weaker.

Table 3: Summary of results for three regression models predicting ITU internet Gender Gap Index using using (i) a single online variable; (ii) offline variables and a single online variable; (iii) offline variables. Table shows coefficients for standardized values of the explanatory variables.

	Online Model			Online-Offline Model			Offline Model
	FB	ADW	FB &ADW	FB & offline var	ADW & offline var	FB & ADW & offline var	Offline indicators
Intercept	0.934	0.933	0.933	0.934	0.932	0.932	0.932
GDP_capita_PPP_2016					-0.029	-0.024	-0.044
HDI				0.049	0.115	0.101	0.110
Mean_year_schooling_HDI					-0.037	-0.037	
Unemployment ratio							0.021
FB GG (age 18+)	0.079		0.066	0.047		0.027	
ADW GG (age18+)		0.067	0.016		0.041	0.025	
Adjusted R-squared	0.528	0.374	0.521	0.646	0.684	0.695	0.585
Mean Abs. Error	0.045	0.053	0.045	0.041	0.038	0.038	0.047
SMAPE	5.7%	6.6%	5.8%	5.3%	5.2%	5.3%	6.3%
F-statistics	92.7	46.4	42.3	76.0	41.0	34.8	38.6
Df	81	75	74	80	70	69	77
N	83	77	77	83	75	75	81

Table 4: Summary of results for three regression models predicting ITU internet Gender Gap Index using using (i) multiple age groups of online variables; (ii) offline variables and multiple age groups of online variables; (iii) offline variables. All reported coefficients are with standardized values of the predictor variables.

	Online Model			Online-Offline Model			Offline Model
	FB	ADW	FB & ADW	FB & offline var	ADW & offline var	FB & ADW & offline var	Offline indicators
Intercept	0.934	0.931	0.931	0.934	0.931	0.981	0.932
HDI				0.038	0.061	0.051	0.110
GDP_capita_PPP_2016							-0.044
Unemployment ratio							0.021
FB GG (age 18+)			0.065				
FB GG (age 20-64)				0.151			
FB GG (age 25-29)				-0.038			
FB GG (age 25-49)	0.082						
FB GG (age 55-59)	0.120						
FB GG (age 50-54)			0.099				
FB GG (age 50+)						0.110	
FB GG (age 60-64)	-0.134		-0.110	-0.073		-0.108	
ADW GG (age 18-24)		0.077	0.024		0.047	0.045	
Adjusted R-squared	0.677	0.473	0.717	0.724	0.696	0.7958	0.585
Mean Abs. Error	0.042	0.051	0.039	0.038	0.036	0.032	0.047
SMAPE	5.3%	6.3%	5.1%	5.0%	4.7%	4.3%	6.3%
F-statistics	58.4	67.3	47.9	54.8	85.8	73.1	38.6
Df	79	73	70	78	72	70	77
N	83	75	75	83	75	75	81

Models Predicting Internet Use Gender Gap

In this section, we present results from three types of regression models, the offline-model, the online-model and the online-offline model. We also report the results after improving the predictive performance of the online and the offline-online models using Facebook and AdWords' estimates for different age groups. All predictor variables were standardized before fitting the model so to make the coefficients of the regression models more comparable to each other.

Offline model: We report the offline model predictive performance summary in Table 3. Various offline indicators were selected by the greedy step-wise forward approach to give an Adjusted R-squared value of 0.58. These variables are the country's GDP per capita, HDI, and the unemployment ratio. Countries that have higher HDI also have higher gender equality in internet use, while countries that have lower unemployment ratios tend to have higher gender equality in internet use. It is interesting to note that even though we included economic and cultural variables as of-

fline indicators, the economic variables are the ones that are picked up in both the offline and online-offline models (reported later). This indicates that economic factors linked to overall country-level development (e.g GDP per capita and HDI) appear to be more important for explaining internet use gender gaps.

Online model: We first estimated three iterations of the online model using the FB GGI for age group 18+, the ADW GGI for age group 18+ and then using both the FB GGI for age group 18+ and the ADW GGI for age group 18+ as predictors of the internet GGI. The results of the online models' predictive performance are shown in Table 3. The positive coefficients on the Facebook and AdWords' GGI variables indicate that gender gaps in Facebook's number of users or Adwords' number of impressions in different countries also reflects gender gap in internet use in those countries. The single-variable online model (with intercept) that uses the FB GGI for age group 18+ only as a predictor has the highest Adjusted R-squared value and lowest SMAPE of the three online models. Comparing with the offline model, the online model that uses the FB GGI for age group 18+ as a predictor generates more accurate out-of-sample estimates with lower Mean Abs. Error and SMAPE.

We further investigated if Facebook and AdWords' indicators for different age groups, either separately or together, could be used to improve the performance of the online model. The findings are presented in Table 4. Several Facebook and AdWords' GGIs for different age groups are positively associated with internet use GGI. Countries with higher FB GGI for age group 25-49 and higher ADW GGI for age 18-24 have higher gender equality in internet use as well. On the other hand, some Facebook GGIs such as the ages 60-64 are negatively associated with internet use GGI. Among all online models, the online model that uses both the FB GGI and the ADW GGI for different age groups shown in Table 4 has the best predictive performance with the highest Adjusted R-squared value and the lowest Mean Abs. Error and SMAPE.

Online-Offline model: We examined if the FB GGI for ages 18+ and ADW GGI for ages 18+ either separately or jointly could be combined with offline predictors to improve predictive performance. The results from these online-offline models are reported in Table 3. Offline variables selected by the step-wise approach include variables linked to economic development such as GDP per capita, HDI and the HDI sub-index for mean years of schooling. The cultural variables are not selected in any of the online-offline models. In general, the online-offline models show better predictive performance than the online models and the offline model. We further examined if additional Facebook and AdWords' indicators for different age groups could be used to improve the performance of online-offline models beyond the global (18+) FB GGI and ADW GGI variables. These findings are displayed in Table 4. On the whole, the online-offline model that uses both the FB GGI and the ADW GGI for different age groups combined with offline indicators performs the best among all the online, online-offline and offline models reported in both Tables 3 and 4. This model produces the highest Adjusted R-squared and the lowest Mean Abs. Error and SMAPE and improves upon the online-offline model reported in Table 3 that uses the FB and ADW GGI for ages 18+ with offline indicators.

Predicting Gender Gaps in ICT skills

As the correlations in Table 2 indicate, the usefulness of the gender gap measures from Facebook and Google advertising extends beyond just predicting the gender gaps in internet use; they could also provide useful proxies for estimating gender gaps in a variety of ICT related skills, which are another important dimension of digital gender gaps. Unlike self-reported indicators in the ITU survey data, these indicators may also capture use of specific platforms, such as Facebook and Google, and thus indirectly capture digital skills of users that are required to make use of these services.

Table 5 provides a summary of regression models providing the best performance that were fitted to predict a variety of different ICT skills gender gaps using a combination of the online FB and ADW GGI and offline variables. The skills shown in the Table highlight a variety of ICT skills ranging from basic editing skills such as using copy and paste tools, file transfer skills to relatively more advanced skills such as numeric skills involving the use of a spreadsheet and computer programming.

For three of the four skill measures, FB GGI measures are selected as the online indicators, with the exception of writing a computer program, for which the ADW measure is selected. The offline indicators selected include development measures such as the Human Development Index (HDI) pertaining to income and education as well as a variable pertaining to gender gaps in holding senior positions in the workplace (Senior managerial GG). Again, as in the internet GGI models, the cultural variables are not selected, indicating again the importance of economic factors in explaining digital gender gaps as well as in improving the predictive power of online variables. For three of the four skills outcomes, online indicators combined with offline indicators provide the models with the best predictive fit. For skills related to transferring files between a computer and a device, online indicators only are picked up.

Some digital skills are more strongly predicted by the variables in the data than others. This can be seen in Table 5 where the models have relatively higher adjusted R-squared and smaller errors for low-level skills such as using copy and paste tools (adj. R-squared 0.727) and file transfer skills (adj. R-squared 0.671) than for high level skills such programming (adj. R-squared 0.325) or using basic arithmetic formula in a spreadsheet (adjusted R-squared 0.505). The coefficients of the predictor variables are positive which indicates that higher FB and ADW GGI are associated with higher gender gap scores across this spectrum of skills. As a result when we observe a smaller gender disparity on Facebook or on Google Adwords we also expect to observe a smaller gender disparity across different types of skills.

A plausible explanation for the differences in the ability of the FB and ADW indicators to capture low-level rather than high-level skills is the types of skills required to use these platforms. For example someone using Facebook or

Table 5: Summary of results for regression models predicting different ICT skills gender gaps. All reported coefficients are with standardized values of the predictor variables.

	Using copy and paste tools	Using basic arithmetic formula in spreadsheet	Writing computer program using programming language	Transfering files between a computer and devices
Intercept	0.903	0.860	0.472	0.851
FB GG (age 15-19)		0.069		0.042
FB GG (age 18-23)	0.100			
FB GG (age 50-54)				0.058
ADW GG (age 25+)			0.091	
Senior managerial work GG		0.038		
Income HDI	0.034			
Secondary Educ. rate HDI			-0.150	
Adjusted R-squared	0.727	0.505	0.325	0.671
Mean Abs. Error	0.056	0.073	0.122	0.054
SMAPE	7.81%	9.33%	28.07%	7.41%
F-statistics	42.31	21.38	11.85	47.85
N	32	41	46	47
Df	29	38	43	44

Google may need the ability to perform tasks such as copy and pasting a search term on Google or moving files around such as when uploading an image on Facebook. As a result we see strong correlations between these types of skills and the gender gap on Facebook than we see for more high level skills such as programming. The fact that FB GGI indicators perform better than ADW GGI indicators for predicting both the internet use and low-level skills could arise from the fact that FB indicators provide counts of users rather than impressions, which is what is provided by AdWords. Being a Facebook user might be a good proxy for basic internet use and digital skills, but as ADW captures a range of different platforms, more active use of these platforms, for example, by male users could make this a less effective proxy for general internet use.

Figures 2 and 3 show the gender gap measure for using copy and paste tools from the data and its value as predicted by the model in Table 5 respectively. As can be seen, for countries where data are available, the model predictions largely agree with the data; however, the coverage of countries is greatly enhanced especially for Africa where data are most lacking.

Discussion

Women's equal participation in the digital society is considered integral to achieve global goals related to gender equality. The ITU has highlighted the importance of "putting in place data, monitoring and evaluation tools around gender equality and ICT, including for measurement of access and use" to realize these goals (International Telecommunication Union, 2015). This study explores how anonymous, aggregate big data from Google and Facebook can help with this endeavor.

Our study contributes to prior research in Fatehkia et al. (2018) that has used Facebook's advertisement audience estimates to predict gender gap in internet use in several ways. First, we evaluated the potential of another novel data source – Google's advertisement impression estimates (AdWords) – to predict gender gaps in internet use around the world. Second, we explored whether Facebook and AdWords' measures for different age groups could be combined to improve models for predicting gender gaps in internet use. Third, we use the latest ground truth ITU data to compute gender gaps in specific types of ICT skills and examine their relationship with Facebook and AdWords' derived measures.

The prediction results are very promising as the online model using Facebook and AdWords' gender gap measures for different age groups is able to explain 72% of the variance in the ground truth of the Internet Gender Gap Index, showing a slight improvement over the online model reported in Fatehkia et al. (2018) in which 69% of the variance was explained.[9] By comparison, the offline model only explains 59% of the variance. Furthermore, the online-offline model that uses *both* the Facebook and AdWords' gender gap measures for different age groups *combined with* offline indicators has the best model fit, explaining around 80% of the variance in the ground truth. This supports our approach to integrate both Facebook and AdWords' gender gap measures to the regression model as using offline measures solely, such as general development and gender-specific development indicators might not be enough to provide good prediction of internet use gender gap. In addition, our approach shows that the predictive performance improves with combining Facebook and AdWords' gender gap measures for different age groups.

Going beyond modeling mere binary internet use-or-not, our results also demonstrate a strong relationship between gender gaps in ICT *skills* and gender gaps in Facebook and Adwords' estimates. Specifically, we found that gender gaps in ICT skills such as sending e-mail with attached files or using copy and paste tools are associated with a gender gap in Facebook and Adwords' estimates. In general, we found

[9]Note that the online model with Facebook GGI 18+ performs worse with the latest round of ITU (2018) used in this study compared with ITU (2016) data used in Fatehkia et al. (2018).

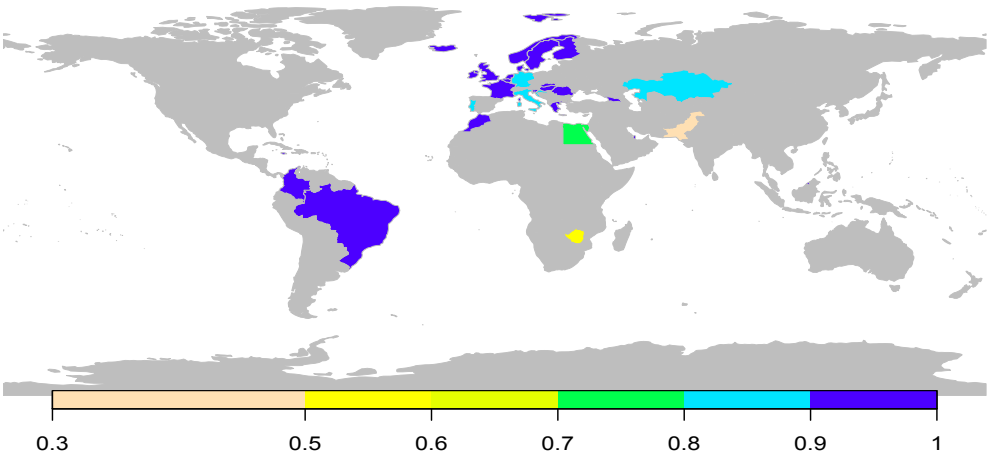

Figure 2: DS GGI-Using copy and paste tools

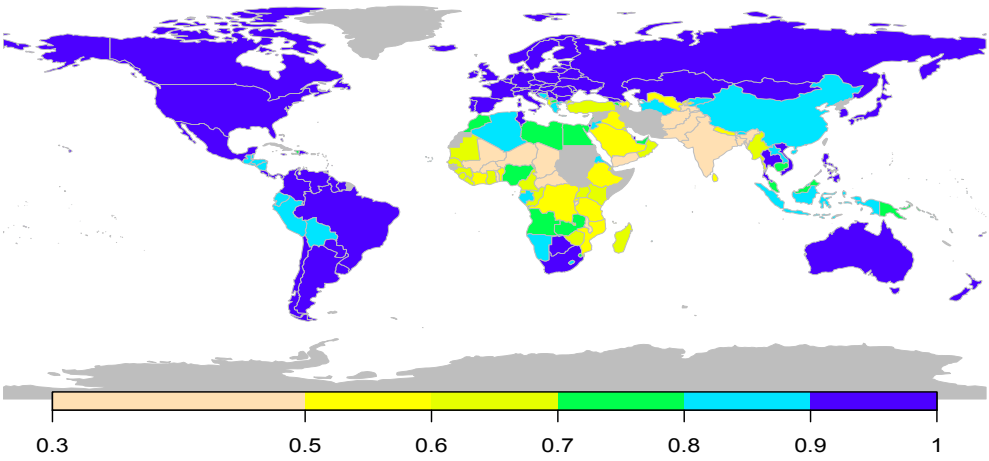

Figure 3: Predicted DS GGI for Using copy and paste tools

that Facebook variables are better able to predict low-level skills, which could be linked to both the type of digital use having a Facebook account is able to capture, as well as the quantity provided by the Facebook marketing platform compared with AdWords (users versus impressions). Further research work might try to use Facebook and AdWords' data to measure gender gaps in digital skills by filtering advertising reach estimates based on education or specific audience's interests.

Using aggregate, anonymous online advertising audience estimates to predict digital gender gaps has several advantages over traditional approaches. First, data from online advertising platforms can be collected regularly, enabling to make predictions on, say, a monthly basis. Second, internet giants such as the Facebook and Google AdWords have a global reach and their data can be collected for more than 190 countries, enabling predictions of internet use and digital skills gender gaps for most countries in the world. Here, the biggest gain in coverage is for low- and lower-middle-income countries where online data enable us to generate estimates for 64 countries, compared with 16 in the ITU data. Third, the Facebook and AdWords' gender gap can be disaggregated by different socio-demographic characteristics such as age, gender, language and location, which is particularly useful for large countries like India. Leveraging these additional characteristics provide further avenues

to extend this work.

Still, our approach of combining Facebook and AdWords' gender gap measures to predict gender gap in internet use has several limitations. First, Facebook and Google do not provide documentation that explains how their algorithms estimate the number of users or the number of impressions. These estimates are hence sensitive to changes in the design of the blackbox algorithms. Second, no online advertising platform that we are aware of provides *historic* estimates, disaggregated by period, for the reach of a to-be-launched ad campaign. This makes it difficult to evaluate changes in the model fit over time. Indeed, we found that the Facebook and AdWords' estimate have a higher correlation with the 2016 ITU data than with the 2018 ITU data. Also, the 2018 ITU data is based on gender-disaggregated data on internet use from 2013-17, depending on the country, whereas the Facebook and the AdWords' data are more recent. As Facebook and Google do not provide historic data, we need to continue collecting data to allow for tracking of change over time and to update our regression models. Nevertheless, our approach of combining Facebook and AdWords' gender gap measures to predict digital gender gaps provides up-to-date monitoring of progress to achieve progress on SDG targets linked to gender equality and digital literacy.

Finally, to extend our approach to monitor other SDGs and other aspects of human development, one always has to keep in mind that not everyone is on the internet, let alone on Facebook or Google. Whereas for this particular study the *absence* of users was the main signal used, in other studies it would be important to complement online data with traditional data from household surveys and censuses to avoid the risk of excluding parts of the population.

Acknowledgements

The research for this paper was conducted as a part of the project 'The Digital Traces for the Gender Digital Divide based at the University of Oxford that received funding from Data2X, an initiative of the United Nations Foundation (Grant No. UNF-17-936).

References

Emad Abu-Shanab and Nebal Al-Jamal. 2015. Exploring the Gender Digital Divide in Jordan. *Gender, Technology and Development* 19, 1 (March 2015), 91–113. https://doi.org/10.1177/0971852414563201

Amy Antonio and David Tuffley. 2014. The Gender Digital Divide in Developing Countries. *Future Internet* 6 (10 2014), 673–687.

Matheus Araújo, Yelena Mejova, Ingmar Weber, and Fabrício Benevenuto. 2017. Using Facebook Ads Audiences for Global Lifestyle Disease Surveillance: Promises and Limitations. In *WebSci*. 253–257.

Joshua Blumenstock, Gabriel Cadamuro, and Robert On. 2015. Predicting poverty and wealth from mobile phone metadata. *Science* 350, 6264 (2015), 1073–1076.

Broadband Commission. 2013. *Doubling Digital Opportunities - Enhancing the Inclusion of Women and Girls in the Information Society*. Technical Report. UNESCO; ITU. http://www.broadbandcommission.org/Documents/working-groups/bb-doubling-digital-2013.pdf

Hyunyoung Choi and Hal Varian. 2012. Predicting the present with Google Trends. *Economic Record* 88 (2012), 2–9.

Rumi Chunara, Lindsay Bouton, John W. Ayers, and John S. Brownstein. 2013. Assessing the Online Social Environment for Surveillance of Obesity Prevalence. *PLOS ONE* 8, 4 (04 2013), 1–8.

Enrico di Bella, Lucia Leporatti, and Filomena Maggino. 2016. Big Data and Social Indicators: Actual Trends and New Perspectives. *Social Indicators Research* (2016), 1–10.

Paul DiMaggio, Eszter Hargittai, Coral Celeste, and Steven Shafer. 2004. *From Unequal Access to Differentiated Use: A Literature Review and Agenda for Research on Digital Inequality*. Technical Report. Russell Sage Foundation.

Christopher D. Elvidge, Paul C. Sutton, Tilottama Ghosh, Benjamin T. Tuttle, Kimberly E. Baugh, Budhendra Bhaduri, and Edward Bright. 2009. A global poverty map derived from satellite data. *Computers & Geosciences* 35, 8 (2009), 1652 – 1660.

European Parliament. 2018. *The underlying causes of the digital gender and possible solutions for enhanced digital inclusion of women and girls*. Policy PE 604.940. Policy Department for Citizen's Rights and Constitutional Affairs. http://www.europarl.europa.eu/RegData/etudes/STUD/2018/604940/IPOL_STU(2018)604940_EN.pdf.

Masoomali Fatehkia, Ridhi Kashyap, and Ingmar Weber. 2018. Using Facebook ad data to track the global digital gender gap. *World Development* 107 (2018), 189 – 209. https://doi.org/10.1016/j.worlddev.2018.03.007

David Garcia, Yonas Mitike Kassa, Angel Cuevas, Manuel Cebrian, Esteban Moro, Iyad Rahwan, and Ruben Cuevas. 2018. Analyzing gender inequality through large-scale Facebook advertising data. *Proceedings of the National Academy of Sciences* (2018). https://doi.org/10.1073/pnas.1717781115 arXiv:http://www.pnas.org/content/early/2018/06/12/1717781115.full.p

Anita Gurumurthy and Nandini Chami. 2014. *Gender equality in the information society*. Technical Report. IT for Change. http://www.itforchange.net/sites/default/files/Final%20Policy%20Brief%20pdf

Nancy J. Hafkin and Sophia Huyer. 2007. Women and Gender in ICT Statistics and Indicators for Development. *Information Technologies & International Development* 4, 2 (Dec. 2007), pp. 25–41. http://itidjournal.org/index.php/itid/article/view/254

Eszter Hargittai. 2002. Second-Level Digital Divide: Differences in People's Online Skills. *First Monday* 7, 4 (2002).

Eszter Hargittai and Steven Shafer. 2006. Differences in actual and perceived online skills: The role of gender. *Social Science Quarterly* 87, 2 (2006), 432–448.

Martin Hilbert. 2011. Digital gender divide or technologically empowered women in developing countries? A typical case of lies, damned lies, and statistics. *Women's Studies International Forum* 34, 6 (Nov. 2011), 479–489. https://doi.org/10.1016/j.wsif.2011.07.001

Geert Hofstede. 2011. Dimensionalizing cultures: The Hofstede model in context. *Online readings in psychology and culture* 2, 1 (2011), 8.

IEAG. 2014. A World that Counts–Mobilising the Data Revolution for Sustainable Development. http://www.undatarevolution.org/wp-content/uploads/2014/11/A-World-That-Counts.pdf

Intel. 2012. *Women and the Web*. Technical Report. http://www.intel.com/content/www/us/en/technology-in-education/women-in-the-web.html

International Telecommunication Union. 2015. *ACTION PLAN TO CLOSE THE DIGITAL GENDER GAP*. Policy. International Telecommunication Union. https://www.itu.int/en/action/gender-equality/Documents/ActionPlan.pdf.

International Telecommunication Union. 2016. Core list of ICT indicators. Retrieved from the International Telecommunication Union website, https://www.itu.int/en/ITU-D/Statistics/Documents/coreindicators/Core-List-of-Indicators_March2016.pdf.

International Telecommunication Union. 2017. Fast-forward progress Leveraging tech to achieve the global goals.

International Telecommunication Union. 2018. World Telecommunication/ICT Indicators Database 2018. Data retrieved from the International Telecommunication Union website, https://www.itu.int/en/ITU-D/Statistics/Pages/publications/wtid.aspx.

IUSSP. 2015. The IUSSP on a Data Revolution for Development. *Population and Development Review* 41, 1 (2015), 172–177. https://doi.org/10.1111/j.1728-4457.2015.00041.x

Emmanuel Letouze and Johannes Jutting. 2014. *Official Statistics, Big data and human development: towards a new conceptual and operational approach*. Technical Report. Data Pop Alliance and PARIS21. https://www.odi.org/sites/odi.org.uk/files/odi-assets/events-documents/5161.pdf

Huina Mao, Xin Shuai, Yong-Yeol Ahn, and Johan Bollen. 2015. Quantifying socio-economic indicators in developing countries from mobile phone communication data: applications to Côte d'Ivoire. *EPJ Data Science* 4, 1 (2015).

Yelena Mejova, Harsh Rajiv Gandhi, Tejas Jivanbhai Rafaliya, Mayank Rameshbhai Sitapara, Ridhi Kashyap, and Ingmar Weber. 2018. Measuring Subnational Digital Gender Inequality in India through Gender Gaps in Facebook Use. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*. ACM, 43.

Sudhakar V Nuti, Brian Wayda, Isuru Ranasinghe, Sisi Wang, Rachel P Dreyer, Serene I Chen, and Karthik Murugiah. 2014. The use of google trends in health care research: a systematic review. *PloS one* 9, 10 (2014), e109583.

Hiroshi Ono and Madeline Zavodny. 2007. Digital inequality: A five country comparison using microdata. *Social Science Research* 36, 3 (Sept. 2007), 1135–1155. https://doi.org/10.1016/j.ssresearch.2006.09.001

Francesco Rampazzo, Emilio Zagheni, Ingmar Weber, Maria Rita Testa, and Francesco Billari. 2018. Mater certa est, pater numquam: What can Facebook Advertising Data Tell Us about Male Fertility Rates? *International Conference on Web and Social Media (ICWSM)* (2018).

Giuliano Resce and Diana Maynard. 2018. What matters most to people around the world? Retrieving Better Life Index priorities on Twitter. *Technological Forecasting and Social Change* (2018).

Laura Robinson, Shelia R. Cotten, Hiroshi Ono, Anabel Quan-Haase, Gustavo Mesch, Wenhong Chen, Jeremy Schulz, Timothy M. Hale, and Michael J. Stern. 2015. Digital inequalities and why they matter. *Information, Communication & Society* 18, 5 (2015), 569–582. https://doi.org/10.1080/1369118X.2015.1012532

Matthew J Salganik. 2017. *Bit by bit: social research in the digital age*. Princeton University Press.

Anique Scheerder, Alexander van Deursen, and Jan van Dijk. 2017. Determinants of Internet Skills, Uses and Outcomes. A Systematic Review of the Second- and Third-level Digital Divide. *Telemat. Inf.* 34, 8 (Dec. 2017), 1607–1624. https://doi.org/10.1016/j.tele.2017.07.007

United Nations Population Division. 2017. World Population Prospects: The 2015 Revision. https://esa.un.org/unpd/wpp/Download/Standard/Population/ https://esa.un.org/unpd/wpp/Download/Standard/Population/.

World Economic Forum. 2016. *Global Gender Gap Report 2016*. Technical Report. World Economic Forum. http://wef.ch/1yrt8iq http://wef.ch/1yrt8iq.

WWW Foundation. 2015. Women's Rights Online: Translating Access into Empowerment. https://webfoundation.org/research/womens-rights-online-2015/

Emilio Zagheni, Ingmar Weber, and Krishna Gummadi. 2017. Leveraging Facebook's Advertising Platform to

Monitor Stocks of Migrants. *Population and Development Review* 43, 4 (2017), 721–734. https://doi.org/10.1111/padr.12102