World Development 107 (2018) 189–209

Contents lists available at ScienceDirect

World Development

journal homepage: www.elsevier.com/locate/worlddev

Using Facebook ad data to track the global digital gender gap

Masoomali Fatehkia [a], Ridhi Kashyap [b,*], Ingmar Weber [c]

[a] Princeton University, USA
[b] University of Oxford, UK
[c] Qatar Computing Research Institute, Qatar

ARTICLE INFO

Article history:
Accepted 3 March 2018

Keywords:
Gender inequality
Internet
Mobile phones
Global digital gender gaps
Big data
Development indicators

ABSTRACT

Gender equality in access to the internet and mobile phones has become increasingly recognised as a development goal. Monitoring progress towards this goal however is challenging due to the limited availability of gender-disaggregated data, particularly in low-income countries. In this data sparse context, we examine the potential of a source of digital trace 'big data' – Facebook's advertisement audience estimates – that provides aggregate data on Facebook users by demographic characteristics covering the platform's over 2 billion users to measure and 'nowcast' digital gender gaps. We generate a unique country-level dataset combining 'online' indicators of Facebook users by gender, age and device type, 'offline' indicators related to a country's overall development and gender gaps, and official data on gender gaps in internet and mobile access where available. Using this dataset, we predict internet and mobile phone gender gaps from official data using online indicators, as well as online and offline indicators. We find that the online Facebook gender gap indicators are highly correlated with official statistics on internet and mobile phone gender gaps. For internet gender gaps, models using Facebook data do better than those using offline indicators alone. Models combining online and offline variables however have the highest predictive power. Our approach demonstrates the feasibility of using Facebook data for real-time tracking of digital gender gaps. It enables us to improve geographical coverage for an important development indicator, with the biggest gains made for low-income countries for which existing data are most limited.

© 2018 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

The rapid proliferation of information and communication technologies (ICTs) has been one of the most significant social phenomena of the new millennium. Today, there are estimated to be more than 7 billion mobile subscriptions worldwide, up from 738 million in 2000. Globally, the number of internet users has risen from roughly 400 million in 2000 to 3.2 billion people who are using the Internet today. 2 billion of these users live in developing countries (ITU, 2015). The expansion of ICTs has paved the way for the digital revolution that is shaping the ways in which people and communities learn and access information and skills, connect with each other, express their ideas, and conduct their everyday lives.

The tremendous potential of digital technologies as tools for empowering communities in poor countries and delivering development goals has become widely recognized among development practitioners (Qiang, Clarke, & Halewood, 2006; Unwin, 2009; Walsham & Sahay, 2006). By lowering the costs of information and connectivity, digital technologies can improve employment opportunities, provide cost-effective health services, and enable access to learning, skills and financial services to help achieve sustainable development goals, particularly for marginalized groups such as women (Broadband Commission, 2013; Hafkin & Huyer, 2006; Huyer & Carr, 2002; Santosham & Lindsey, 2015; WWW Foundation, 2015). This commitment is explicitly noted in the United Nations (UN) Sustainable Development Goals (SDGs), in which Goal 5(b) pledges to "enhance the use of ...information and communication technology to promote the empowerment of women".[1]

Although significant reductions in the costs of ICTs have facilitated their rapid expansion in the developing world in the past two decades, significant gender inequalities in access to these technologies persist. The International Telecommunications Union (ITU), the UN's specialized agency for ICTs, estimates that some 200 million fewer women are online compared with men (Broadband Commission, 2013). Although the digital divide, a term used to describe inequalities in access to the internet, between

* Corresponding author at: Department of Sociology/Anthropology/Nuffield College, Oxford OX1 1NF, UK.
E-mail address: ridhi.kashyap@nuffield.ox.ac.uk (R. Kashyap).

[1] https://sustainabledevelopment.un.org/post2015/transformingourworld.

https://doi.org/10.1016/j.worlddev.2018.03.007
0305-750X/© 2018 The Author(s). Published by Elsevier Ltd.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

men and men is not restricted to developing countries, gender inequalities in access are greater in the developing world (Alozie & Akpan-Obong, 2017; Broadband Commission, 2013; Hafkin & Huyer, 2007). As a result, concerns about gender inequalities in internet and mobile access have emerged as an important focus of the discussions on ICTs for development (Broadband Commission, 2013; Moolman, Primo, & Shackleton, 2007; Santosham & Lindsey, 2015). In March 2013, the International Telecommunications Union (ITU) and UNESCO Broadband Commission for Digital Development further endorsed a target calling for gender equality in access to broadband by 2020 (Broadband Commission, 2013).

Despite the increasing visibility of the issue, our ability to measure gender gaps in access to the internet and mobile phones is significantly limited due to data gaps. A report by the ITU/UNESCO identified the lack of sex-disaggregated data on internet and mobile phone access as "one of the key barriers" for measuring progress in development goals that call for gender equality in access to the internet (Broadband Commission, 2013, p. 19). Official, nationally representative gender-disaggregated statistics on internet and mobile access lack coverage across countries and regular production, and data availability on these indicators is especially limited in low-income countries (Antonio & Tuffley, 2014; Brännström, 2012; Hafkin & Huyer, 2007). When data are available, inconsistencies across data sources and measures impede cross-national comparisons. Furthermore, these data lack granularities at the subnational level or by socio-demographic group, such as age or education.

In this data sparse context, there is the need to look at other data sources that can be used to measure and track the current state of digital gender disparities. The continually updating nature of 'big data' sources, particularly digital trace data, that capture online footprints left behind on digital spaces in real time offer promise in this setting for nowcasting, or in other words generating real-time predictions of social outcome indicators in the present (di Bella, Leporatti, & Maggino, 2018). 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" (Blumenstock, Cadamuro, & On, 2015; Mao, Shuai, Ahn, & Bollen, 2015; Elvidge et al., 2009; Evangelos, Efthimios, & Konstantinos, 2013; Giannone, Reichlin, & Small, 2008; Harald et al., 2013; Lampos & Cristianini, 2012; Yazdani & Manovich, 2015).

This paper leverages one such source of digital trace data – Facebook's advertisement audience estimates – to measure and nowcast digital gender gaps in internet and mobile access in a global perspective. We generate a unique country-level dataset combining 'online' indicators of Facebook users by gender, age and device type, 'offline' indicators related to a country's overall development and gender-specific development indicators, and official data on gender gaps in internet and mobile access where available. As not all those with access to the internet and mobile phones are Facebook users, we need to examine the validity of Facebook-derived data for measuring digital gender gaps more broadly. To do this, we predict gender gaps in internet and mobile phone access from official data using online, Facebook-derived indicators, as well as models combining online and offline indicators.

Our results demonstrate the feasibility of using Facebook data to measure digital gender gaps around the world. Models using only online Facebook data do better than models using only offline indicators for predicting internet gender gaps as reported in official ITU statistics. Models combining Facebook data with offline gender equality indicators however do the best. With the online model using Facebook data, we estimate a mean value of the internet gender gap index for low-income countries to be 0.76. This implies that 0.76 women have access to the internet for every man who does, indicating that female internet penetration is 24% lower than that for males. This index increases to 0.86 for lower-middle income countries and to 0.92 and 0.95 for upper-middle and high income countries respectively based on the World Bank's classification. For the mobile phone gender gaps our online model predicts a gender gap index of 0.79 for low income countries indicating a female deficit in mobile access of 21%. This increases to 0.90 and 0.96 for lower-middle and upper-middle income countries respectively, with high income countries having an index of 0.98.

The Facebook data enable us to improve coverage for indicators of the internet gender gap significantly from 84 countries,[2] for which official data are available from the ITU, to 152. The biggest gains are for low- and lower-middle-income countries, for which the Facebook data enable us to generate measures for 64 countries compared with 13 in the ITU data. For mobile phone gender gaps, the Facebook data enable us to expand coverage from 22 countries to 153, with the biggest gains again for low- and lower-middle-income countries where we are able to generate measures for 64 countries instead of the 17 in currently available sources. Our work presents an example of data innovation that harnesses the potential of 'big data' sources for real-time and global monitoring of development indicators (IEAG, 2014; Letouze & Jutting, 2014).

2. Digital gender gaps

2.1. The data gap: gender and ICTs

The 'digital divide' is a concept that has been used to describe inequalities in access to ICTs, in particular access to and use of the internet (Norris, 2001). These could refer to disparities across different dimensions, for example between developed and developing countries, as well as those between groups within countries, such as those between men and women, or the rich and the poor. In this paper, our focus is on the digital gender gaps, specifically gaps in access to the internet and mobile phones between men and women. Digital gender gaps can be of various forms, including differences in access, the extent of use, in technical skills, and social support in using these technologies (Broadband Commission, 2013; Bimber, 2000). Given significant variation in internet and mobile penetration rates, global analyses of the digital gender gap have predominantly referred to the gaps in access and use of the internet without reference to specific types of uses or skills.

As with any technology, the expansion of ICTs has followed a diffusion process with early adopters and late adopters (Rogers, 2010). Studies from industrialized countries such as the US and Canada have shown that patterns of early adoption online often mirrored social inequalities from the offline world, as high income, educated and urban groups, and men were more likely to go online earlier (DiMaggio, Hargittai, Neuman, & Robinson, 2001). The increasing diffusion of digital technologies in industrialized countries has generally accompanied a reduction in gender gaps in their access (Haight, Quan-Haase, & Corbett, 2014; Ono & Zavodny, 2007; Rice & Katz, 2003).

Although gaps in access have closed, differentiated patterns of use between men and women have emerged (DiMaggio, Hargittai, Celeste, & Shafer, 2004). Men remain more frequent users of the internet, reporting higher levels of daily usage, and are more likely to use the internet for activities such as games, music and online trading, whereas women were more likely to use the internet for instant messaging and staying connected

[2] Of these 84, we obtained Facebook ad audience data for 78 countries. Three countries, Iran, Cuba and Sudan, are not currently supported by Facebook. Three other countries, Montserrat, Palestine and Puerto Rico, were not supported in April 2017 when we compiled the list of countries to consider.

(Dunahee & Lebo, 2016; Ono & Zavodny, 2007; Wasserman & Richmond-Abbott, 2005). The Pew Research Centre's Internet and American Life Surveys found that women were more likely to be on social media websites such as Facebook in the US, although by 2015 the gap between men and women's social media had closed (Pew Research Centre, 2015, 2016). Studies based on samples from industrialized countries such as the US and Germany have found that women and men use social media differently. Women tended to use social media to maintain social ties and for self-presentation such as by posting photographs, whereas men were more likely to use social media to acquire general information (Haferkamp, Eimler, Papadakis, & Kruck, 2012; Joinson, 2008; Krasnova, Veltri, Eling, & Buxmann, 2017). In their review of the literature on social network sites, Boyd and Ellison concluded "scholars still have a limited understanding of who is and who is not using these sites, why, and for what purposes, especially outside the U.S" (Boyd & Ellison, 2010, p. 224).

Among low-income countries, a handful of field research studies have noted significant gender gaps in ICT access, but systematic and regularly available quantitative data on gender-disaggregated ICT use have remained relatively limited (Broadband Commission, 2013; Intel, 2012; Santosham & Lindsey, 2015; WWW Foundation, 2015). The ITU has collected gender-disaggregated data since 2007 through an annual questionnaire that it sends to member states (Hafkin & Huyer, 2007). These surveys seek to collect data on internet use by gender of the user. Although it is sent to all member states, the data on this indicator are largely available for high-income and upper-middle income countries as per the World Bank classification. Low-income and lower-middle income countries, where internet penetration rates are low and gender gaps are likely to be large, generally lack regularly updated indicators on internet or mobile users by gender. For example, most recently available data compiled by the ITU based on these surveys were available for 84 countries, of which 45 were high-income countries, 24 were upper-middle income countries, 11 were lower-middle income and 2 were low-income countries.[3] Data on these indicators from Sub-Saharan Africa (4 out of 48 countries) and South Asia (1 out of 8) are especially limited.

Although ITU compiled data provide indicators for internet users by gender, gender-disaggregated statistics on mobile phone usage are rarer (Santosham & Lindsey, 2015). Telecommunications authorities collect national aggregate data on telephone subscriptions, devices and connections, but data by individual attributes are not usually collected (Hafkin & Huyer, 2007). Individual mobile phone operators do not systematically collect data based on individual attributes, but in some cases these data are available through SIM ownership records from mobile phone operators. When available, however, this information is not widely disseminated for public use, nor is it used to inform government statistics and policy. Population censuses provide another potential source for ICT data. Censuses are able to provide household level information on the physical availability of the internet or mobile phones within households, but are not able to capture differences in use of the technology within the household.

Based on data compiled from countries with available survey data, the ITU estimates that while about 45% of men in developing countries were using the internet, 37.5% of women were using the internet in the same (ITU, 2017).[4] For the least developed countries, they estimate male internet penetration to be 21%, compared to female internet penetration of 14%. The female internet penetration rate here is defined as the proportion of women, as a fraction of the female population, who have access to the internet (ITU, 2017). For mobile phones, the GSMA, a trade body representing the interests of mobile operators worldwide, estimates that 59% of women in the world do not own mobile phones whereas 74% of men do. In absolute terms, this equals 80 million fewer females than males in low- and middle-income countries who do not own mobile phones, with two-thirds of this population in South Asia where 72% of women do not own mobile phones (Santosham & Lindsey, 2015).

2.2. Predictors of the digital gender gap

Early studies on gender gaps in internet use from industrialized countries attributed it to a combination of socio-economic differences, particularly gender gaps in income, employment and education between men and women (Ono & Zavodny, 2007; Bimber, 2000) as well as lack of confidence resulting from women underestimating their actual usage skills (Cooper, 2006; Hargittai & Shafer, 2006; Joiner, Messer, Littleton, & Light, 1996). Studies from the US and Canada have found that gender gaps have generally disappeared with increasing internet penetration (DiMaggio et al., 2004; Haight et al., 2014), and there is "some support at the macro level that the gender divide moves in the same direction with overall internet penetration" (Hafkin & Huyer, 2007, p. 35). Hafkin and Huyer however caution that this relationship "is tenuous at best" and that the gender divide cannot simply be expected to improve as overall infrastructure for digital technologies improves and they become more widespread (Hafkin & Huyer, 2007, p. 35). For example, based on data from the early 2000s the authors reported that the proportion of female Internet users in the Netherlands (40%) was the same as that in Brazil, Mexico, Zimbabwe and Tunisia, despite the significantly higher levels of internet penetration in the Netherlands.

In a study of gender gaps and ICT use in Latin America and 11 countries in Africa drawing on individual-level data from household surveys, Hilbert (2011) found that gender gaps in both internet and mobile use could be entirely explained by women's lower levels of literacy, employment and incomes within each country. Once these variables were controlled for, women often turned out to be more active users of ICTs. Combining insights from Hilbert's study with those from Hafkin and Huyer, we would expect gender-specific development indicators, such as gender gaps in education, labor market outcomes and income-related measures, to be equally if not more important predictors of cross-country variation in digital gender gaps than indicators of overall level of development or wealth alone. Indicators of offline gender gaps are likely to capture the financial and institutional constraints, such as in limited education and income that women directly face in going online. Existing work has highlighted women's access to education as an especially important correlate of women's internet use in low-income countries (Antonio & Tuffley, 2014; WWW Foundation, 2015). Narrowing gender gaps in education have been linked to other favorable overall as well as gender-specific development outcomes (Abu-Ghaida & Klasen, 2004; Ganguli, Hausmann, & Viarengo, 2014).

Aside from financial or institutional constraints limiting access, cultural factors may also prevent women from going online. The internet and mobile phones enable women to participate in the public sphere. In patriarchal societies, this access may be threatening to men and male-dominant institutions that may seek to control or limit women's access to these technologies. For example, in India, village councils in the state of Uttar Pradesh have sought to ban use of mobile phones by single women,

[3] Based on the World Bank classification, available here: https://data-helpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups. Income classification was not available for Montserrat and Palestine, for which ITU reports data.

[4] The classification of countries as developing, developed or least developed here refers to the UN M49 classification of countries as adopted by the ITU. Country classifications are available here: http://www.itu.int/en/ITU-D/Statistics/Pages/definitions/regions.aspx.

arguing that they encourage 'loose behavior' (Gurumurthy & Chami, 2014). Research from Zambia found that women were threatened, intimidated and even beaten by spouses when they utilized their mobile phones (Gurumurthy & Chami, 2014).

Another type of cultural barrier that may result in digital gender gaps comes from women choosing to avoid going online and seeking new technologies due to a lack of self-belief in their ability to master them. This belief may be reinforced by existing stereotypes in families or educational institutions that may influence how girls and women engage with technology (Abu-Shanab & Al-Jamal, 2015; Antonio & Tuffley, 2014). A study by Intel found that one in five women in India and Egypt believe that the internet is not appropriate for them, or that their families would disapprove of their use of it (Intel, 2012).

These aforementioned barriers in countries with conservative gender norms can lead to surprising selection bias effects among women who are online in these countries. Magno and Weber (2014) found that although in countries such as Egypt or Pakistan far fewer women are on Google+ compared to men, those women who are online in these countries have *higher* status than men as indicated by a higher number of followers and a higher PageRank centrality in the social graph compared to their male counterparts. Conversely, in more gender egalitarian countries such as Norway and Finland, roughly equal numbers of men and women are on Google+, but here the men have a higher status online. Similarly, Messias, Vikatos, and Benevenuto (2017) found that within the US, male Twitter users tend to have more followers than female Twitter users, even though Twitter use was relatively balanced among women and men.

Magno and Weber speculated that the global trend they observed was due to a "Jackie Robinson Effect", similar to the situation where female politicians perform better than male politicians as just performing equally would not suffice to get them to such a position in the first place (Anzia & Berry, 2011). Wagner, Graells-Garrido, Garcia, and Menczer (2016) observed a similar effect on Wikipedia where those women who are covered are more likely to be "notable" compared to their male counterparts.

In sum, we should expect women to be less represented online relative to men in poorer, less developed countries with poorer ICT infrastructure. Digital gender gaps, however, are not attributable to overall development alone, and are also likely to be influenced by broader gender inequalities such as education and the economy.

3. Data

Our study seeks to estimate the magnitude of gender gaps in internet and mobile phone access in a global perspective using online Facebook-derived indicators in combination with offline, country-level development indicators. The online data that we use are Facebook's advertisement audience estimates that come from Facebook's marketing application programming interface (API). These data are publicly-accessible and allow advertisers or any user with a Facebook account to query the number of Facebook users disaggregated by various geographic and demographic attributes, such as user age and gender.[5] The offline data come from different data sources and include indicators for gender-specific internet users from the ITU and mobile users from the GSMA, as well as general development and gender-specific development indicators. In what follows, we describe these two classes of data in more detail.

[5] Information and documentation about Facebook Marketing API is available here: (https://developers.facebook.com/docs/marketing-apis).

3.1. Variable of interest: facebook gender gap index

Facebook makes its revenue from targeted advertising and to support advanced targeting options creates detailed profiles of its users including demographic attributes such as age, gender, education, interests expressed on Facebook and information on the device used to access Facebook. These data are implicitly made available to advertisers who can choose to show their advertisement to, say, men aged 18+ living in India and using an iPhone 7. Before the advertisement is launched and before any cost is incurred, Facebook provides the advertiser with information on the audience size, i.e. on how many users match the given targeting criteria, in this case 470,000.[6] This information is of use to the advertiser as it affects the total expected ad cost.

From a social research perspective, the same data can be used as a kind of digital census across Facebook's roughly 2 billion users which can be queried with questions such as "how many Facebook users match criteria X". As more than half of all internet users are also Facebook users,[7] the data have the potential to capture sizable populations in real-time. These data have been used to study population health (Araújo, Mejova, Weber, & Benevenuto, 2017; Saha, Weber, Birnbaum, & De Choudhury, 2017; Chunara, Bouton, Ayers, & Brownstein, 2013), as well as recently, to derive estimates of stocks of migrants (Zagheni, Weber, & Gummadi, 2017). Zagheni et al. (2017) found that even without bias correction the Facebook data are able to provide good demographic estimates of quantities such as percentages of the population of a particular nationality. These features make this data source particularly appealing to monitor digital gender gaps.

For the purposes of this study, data on the estimated number of Facebook users disaggregated by age and gender as well as the type of device used to access Facebook were retrieved for 193 countries.[8] We focus on country-level estimates as the data against which we validate our Facebook-derived estimates (detailed below) are only available at the country level. The Facebook data can also be collected at the subnational level and also be used to capture other sociodemographic attributes (e.g. education) of the user. The device types for which data were collected are mobile devices (all types), feature phones, smart phones, iOS devices, Android devices and iPhone 7.

These data were then used to compute measures of gender gaps in online presence as seen on Facebook. We define the Facebook Gender Gap Index (FB GGI) as follows:

$$\text{Facebook Gender Gap Index} = \frac{\text{Female to male gender ratio of people with characteristic}}{\text{Female to Male gender ratio of the population}} \quad (1)$$

In Eq. (1), the characteristics could be, for example, Facebook users aged 18 and older, or Facebook users aged 18 and older who use a particular mobile device (such as a feature phone) to access Facebook. We divide the gender ratio (female to male) of users of a certain characteristic on Facebook with the population gender ratio (of the same age category as the Facebook ratio) in Eq. (1) for two reasons. First, this information allows us correct the Facebook-derived measures for population imbalances. For example, countries such as Qatar 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 Facebook users for Qatar could

[6] As of July 25, 2017.

[7] According to http://www.internetlivestats.com/internet-users/, there are 3.7 billion internet users as of August 16, 2017, compared to 2 billion Facebook users according to https://techcrunch.com/2017/06/27/facebook-2-billion-users/.

[8] The 193 countries come from the list of 195 countries supported by Facebook's marketing API as of April 2017; Netherlands Antilles and Vatican were excluded from this list as they had Facebook counts of 20, which Facebook's API uses as a placeholder for zero counts. The number of countries/territories supported has grown since then to 246 of which 14 return Facebook counts of 20 (as of August 8, 2017).

be merely reflecting the population gender imbalance. Second, this enables us to approximate a gender-specific Facebook penetration measure that is akin to the definitions of gender-specific internet penetration or mobile ownership measures against which we validate the Facebook measures, as described in Section 3.2. We obtained the population gender ratios for various age groups from the UN World Population Prospects Database (United Nations Population Division, 2017).

All gender gap indicators used in this study, such as those defined in Eq. (1) and later in Eq. (2), take values greater than zero, with values less than 1.0 indicating countries where women are doing worse than men. Values close to 1.0 are desirable as they indicate gender parity. In this sense, a higher value on this index indicates a gender gap that has been *closed*. This is in line with the methodology of the Global Gender Gap Report (GGGR) (World Economic Forum, 2016), data from which are also used in our analysis. Also in line with the GGGR methodology, 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).

Data on the number of Facebook users in the 193 countries disaggregated by gender and age were collected on June 21, 2017 and further disaggregated data by device type and gender were collected on July 25, 2017. Data disaggregated by gender, age and device type were again collected on the 15th of each month from September to November 2017 in order to analyze the temporal variations over a period of several months. The data were collected using the pySocialWatcher library, a Python language library that automates the data collection process by making calls to Facebook's marketing API (Araújo et al., 2017).[9] Section D in the appendix describes the Facebook data in further detail. Since Facebook's API never returns an estimated audience of less than 20 users, presumably to protect user privacy through the concept of k-anonymity (Sweeny, 2002), counts of 20 were removed.

3.2. Dependent variables: internet and mobile gender gap index

The data from the Facebook marketing API provide us estimates of Facebook users by gender. However, not all internet users are Facebook users nor do all users of mobile phones access Facebook on their phones. In order to evaluate the extent to which gender gaps in Facebook use capture gender disparities in internet and mobile use more generally, we need to compare Facebook estimates with those from reliable 'ground truth' measures of internet and mobile use by gender.

As highlighted in Section 2.1, data on ICT use by gender are often limited and not regularly updated, particularly for low- and lower-middle-income countries. The best, currently available database for internet use by gender of the user is compiled by the ITU (ITU, 2016). These data provide proportions of individuals using the internet by gender of the user 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. As the surveys were fielded in different years in the different member states, most recently available estimates vary between 2011 and 2015.

Following the methodology of the Global Gender Gap Report (GGGR), gender gap indices are generally defined as:

$$\text{Gender Gap Index} = \frac{\%\text{ of female population with characteristic}}{\%\text{ of male population with characteristic}} \tag{2}$$

Eq. (2) can be used to generate an internet gender gap index (Internet GGI) using the ITU data on gender-specific internet penetration rates. As with all GGIs, a value of 1 in the Internet GGI indicates a completely closed gender gap where male and female internet penetration rates are equal. The ITU Internet GGI serves as the ground truth measure against which we seek to validate our Facebook-derived gender gap measures. The GGGR itself does not include an ICT-related indicator.

For mobile phone gender gaps, similar data from the ITU are not available. Instead we use data compiled by the GSMA that draws on information from mobile phone companies as well the GSMA's own field research to measure male and female unique subscriber penetration rates, which capture, similarly to the internet indicator, the proportion of men or women in a country as a fraction of all men and women respectively that own a mobile phone. These data are available for 22 countries (GSMA Intelligence, 2015a, p. 17).[10] From the GSMA data, we compute the GSMA Mobile Gender Gap Index (Mobile GGI). The ITU Internet GGI and GSMA Mobile GGI are the two dependent variables against which we validate the potential of the Facebook-derived indicators for predicting gender gaps in internet and mobile access.

3.3. Other predictor variables

Our key predictor variable of interest is the Facebook GGI described in Section 3.1 that comes from online data. Other offline indicators are also likely to be associated with digital gender gaps, and are likely to help us predict our two ground truth measures of digital gender gaps, the ITU Internet GGI and GSMA Mobile GGI. Based on literature summarised in Section 2.2, these include indicators related to both general development and ICT infrastructure in a country as well as more gender-specific development indicators that capture different domains of gender inequalities in a country.

General development-related indicators that are included in our dataset are the Oxford Multidimensional Poverty Index (OMPI) (Alkire & Robles, 2017) and log GDP per capita in 2016 (World Bank, 2017), whereas ICT infrastructure related variables included are internet penetration (Miniwatts Marketing Group, 2017) and unique mobile phone subscriber penetration (USP) rates (GSMA Intelligence, 2016a,b, 2017a,b, 2015b). Internet penetration is defined according to Internet World Stats (IWS) as the number of internet users as a percentage of the total population where an internet user is defined as someone with access to an internet connection and the knowledge to use the web (Miniwatts Marketing Group, 2017). The USP is defined by GSMA Intelligence as the number of unique cellular subscribers as a fraction of the total population (GSMA Intelligence, 2015a, p. 3). This indicator variable takes the value of 1.0 if the USP is below 40% and a value of 0.0 otherwise. Our coding of the variable differs slightly from the definition used in the GSMA Intelligence report (GSMA Intelligence, 2015a, p. 15) where this same variable is used in a regression model to predict the mobile phone gender gap for countries lacking these data. There the indicator variable takes a value of 1.0 if the USP is below 25%. The reason for this change in definition was that none of the

[9] The library package is available here: https://github.com/maraujo/pySocialWatcher.

[10] For the case of mobile phone gender gap index, the GSMA ground truth data are reported as the difference in the male and female subscriber penetration rates divided by the male subscriber penetration rate (GSMA Intelligence, 2015a). This is equivalent to (1 – Gender Gap Index). The GSMA data were transformed to Eq. 2 to have a consistent definition of gender gap throughout. As this is a linear transformation it does not affect the correlations and regression models, apart from a multiplicative factor of −1 and a constant offset. No country had a mobile phone access gender gap index greater than 1.0 and so the truncation at 1.0 did not affect the results.

Table 1
Summary statistics of the variables used in the analysis. Q1 and Q3 are the first and third quartiles respectively. Std is the standard deviation.

Variable	Min	Q1	Median	Mean	Q3	Max	Std.
Dependent (Ground Truth) Variables							
ITU Internet GGI	0.529	0.910	0.966	0.934	0.984	1.000	0.084
GSMA Mobile GGI	0.490	0.700	0.865	0.820	0.948	0.990	0.157
FB Predictor Variables							
FB GGI age 18+	0.164	0.546	0.827	0.749	0.992	1.000	0.244
FB GGI age 25–29	0.126	0.556	0.850	0.765	0.990	1.000	0.245
FB iOS device GGI	0.251	0.544	0.904	0.783	1.000	1.000	0.242
Other (Offline) Predictor Variables							
log(GDP per Capita)	6.549	8.293	9.399	9.267	10.190	11.760	1.199
Internet Penetration	0.013	0.277	0.532	0.522	0.771	1.006	0.288
GGGR Score	0.516	0.666	0.699	0.697	0.729	0.874	0.056
GGGR – Literacy	0.493	0.937	0.989	0.928	1.000	1.000	0.123
GGGR – Education	0.618	0.959	0.991	0.960	0.999	1.000	0.072
GGGR – Tertiary Educ.	0.197	0.909	1.000	0.907	1.000	1.000	0.176
GGGR – Economy	0.320	0.608	0.670	0.658	0.733	0.865	0.113
OMPI	0.000	0.018	0.101	0.160	0.282	0.605	0.166
low USP Indicator	0.000	0.000	0.000	0.160	0.000	1.000	0.368

countries in our dataset for the most recent period had a unique subscriber penetration below 25% as subscriber penetration rates have increased over time.

For measures of gender gaps at the country-level we used data from the World Economic Forum's Global Gender Gap Report (GGGR) (World Economic Forum, 2016) which measures gender equality along four main dimensions – economic, health, political and educational equality – each derived from a set of sub-indices. The GGGR Education variable is a sub-index of gender gaps in educational attainment and is computed as a weighted average of the female to male ratios for primary, secondary and tertiary education as well as the ratio of the female literacy rate to the male literacy rate (World Economic Forum, 2016, p. 5). The GGGR Literacy variable is the ratio of the female literacy rate to the male literacy rate. The literacy rate for each gender is the percentage of the population of that gender "aged 15 years and over who can both read and write and understand a short simple statement on his/her everyday life" (World Economic Forum, 2016, p. 73). The GGGR Tertiary Education Enrollment variable measures the female to male ratio of the total tertiary enrollments expressed as a percentage of the five year age group beyond the official secondary school graduation age (World Economic Forum, 2016, p. 73).

Summary statistics of the dependent variables, the Facebook variables and other predictor variables are reported in Table 1. Table 2 lists the variables used in our models, including their features (geographical coverage, latency or frequency of data generation) that are important to consider for nowcasting, and their correlations with the ground truth (ITU) internet and mobile phone (GSMA) GGI measures. The variables listed in Tables 1 and 2 are those selected in the models that we present in the paper. We also additionally compiled and experimented with other variables in the analysis. A full list of these variables is available in Table 9 in the Appendix.

As shown in Table 2, the number of countries for which each variable was available varied. Where data were not available for a specific country, it was left as a missing field in the data set. Throughout the analysis such as when computing correlations or estimating regression models, the largest set of countries that have all the required data are used for computation. As a result the fitted models differ slightly in the number of observations used. All tables report the number of countries used in the computation.

4. Methods

Our statistical approach involves building regression models that attempt to predict the ground truth measures for digital gender gaps, namely (i) the ITU Internet GGI and (ii) the GSMA mobile GGI using data from the online (Facebook derived) and offline indicators from our dataset. We use ordinary least squares (OLS) regression models for prediction to (i) avoid overfitting as more complex, non-linear models typically require more labeled data for model fitting, and to (ii) have a model that is easily interpretable and can hence be checked for plausibility.

For each of the dependent variables we fitted three different models, namely (i) an online model, (ii) an online-offline model and (iii) an offline model. The online model is a parsimonious, single variable model fitted using the most predictive Facebook GGI variable. Due to its reliance on Facebook indicators as predictor variables, this model has low latency and wide geographical coverage in comparison to other predictor variables as shown in Table 2. The latter two models allow us to compare between online and offline indicators in their predictive potential for internet and mobile phone gender gaps. The online-offline and offline models[11] were arrived at using a greedy step-wise 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. For the three models we also performed greedy step-wise forward model selection with a fivefold cross validation approach whereby the dataset was split into five portions (Friedman, Hastie, & Tibshirani, 2001). Each time a single portion was left out as a validation set and the step-wise forward model selection was performed on the remaining portion with variables being added if the mean squared error of the resulting model on the validation set was the lowest. This resulted in five slightly different models allowing us to compare our models further for out-of-sample robustness and predictive power.

To evaluate the performance of different regression models for predicting digital gender gaps, we report four measures of model fit: (i) Adjusted R-squared, (ii) Pearson r correlation, (iii) Mean Absolute Error, and (iv) Symmetric Mean Absolute Percentage Error (SMAPE). For the first two, 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 one entry, and the fitted model is then used to predict the left out data point. This last error metric is a measure of out-of-sample prediction error indicative of how well the model can predict future unseen observations.

[11] An exception is the offline model for the mobile phone gender gaps where we replicated the model appearing in the report by GSMA Intelligence where the mobile gender gap data come from (GSMA Intelligence, 2015a).

Table 2
Data availability, features, and correlations of the different variables used in the analysis with the ITU Internet Gender Gap Index and GSMA Mobile Gender Gap Index. The number in parenthesis indicates the number of countries for which the correlation is computed based on data availability of the indicator.

Variable	Number of Countries with data	Correlation with ITU Internet GGI	Correlation with GSMA Mobile Phone GGI	Latency/(Latest year of data)
ITU Internet GGI	78	1.00 (78)	.820 (7)	varies (2011–15)
GSMA Mobile Phone GGI	22	.820 (7)	1.00 (22)	varies (2015)
Internet Penetration	191	.650 (78)	.564 (22)	Annual (2016–17)
log(GDP per Capita)	175	.592 (74)	.593 (22)	Annual (2016)
OMPI	100	−.765 (22)	−.572 (22)	Biannual (2005–15)
low USP Indicator	100	−.581 (35)	−.767 (21)	Annual (2015–16)
South Asia Indicator	191	−.425 (78)	−.613 (22)	Static (2017)
GGGR Score	139	.297 (75)	.079 (19)	Annual (2015)
GGGR – Economy	139	.380 (75)	.330 (19)	Annual (2015)
GGGR – Education	139	.618 (75)	.241 (19)	Annual (2015)
GGGR – Literacy	138	.510 (74)	.348 (19)	Annual (2015)
GGGR – Tertiary Educ.	132	.642 (73)	.141 (19)	Annual (2015)
FB GGI age 18+	178	.834 (76)	.650 (22)	Real time (2017)
FB GGI age 25–29	179	.808 (77)	.739 (22)	Real time (2017)
FB iOS device GGI	178	.815 (76)	.481 (22)	Real time (2017)

5. Results

5.1. Correlation analysis with internet and mobile phone gender gaps

Table 2 presents variables in our models and their correlations with the ITU internet and GSMA mobile phone gender gap measures. The Facebook GGI for ages 18+ is the most strongly correlated (Pearson's correlation coefficient of 0.83) variable with the ITU Internet GGI. Similarly, the Facebook GGI for ages 25–29, with a correlation of 0.74, is among the most strongly correlated variables with the GSMA Mobile Phone GGI. These correlations indicate that the Facebook derived measures of gender gap are highly predictive of the ground truth digital gender gap measures.

Among offline indicators, the level of internet penetration has a high correlation (0.65) with the ITU Internet GGI. For the mobile phone gender gap, the indicator variable for low unique subscriber penetration (USP) (−0.77) is important. These results indicate that the level of ICT infrastructure is correlated with gender gaps in ICT access. Women tend to be late adopters, but greater diffusion of technology is accompanied by a closing of the gender gap.

Country-level general development related indicators such as log GDP per capita as well as poverty indicators such as the OMPI have relatively strong correlations with both internet and mobile phone gender gaps. Among the various offline gender gap indicators, GGGR variables for gender gaps in education, literacy, and tertiary education enrollment are the most strongly correlated with internet gender gaps (values above 0.5).

These correlations highlight the interconnection of overall development, ICT infrastructure and gender disparities in other domains such as education and economic opportunities with gender gaps in internet and mobile phone access at the country level. The strongest correlations, at least for internet gender gaps, are with the Facebook measures. This demonstrates the viability of using Facebook measures to predict the internet and mobile phone gender gaps in countries for which data on these indicators are not available.

5.2. Regression models predicting internet gender gap

In this section we present results from three regression models, as described in Section 4, that draw on online, online-offline and offline variables as predictors for the ITU internet GGI. The best-fit model of each of the three categories are summarized in Table 3. All predictor variables were standardized before fitting the model so the coefficients of the regression models can be interpreted as the increase in the value of the Internet GGI associated with an increase of one standard deviation from its mean for the predictor variable. A full list of online and offline variables that we experimented with in our analysis is provided in Table 9 in the Appendix.

The offline model with the highest R-squared, in addition to the internet penetration measure, includes the GGGR gender gap measures for economy and tertiary education as predictors of the internet gender gap. Countries that score higher on gender equality in those domains also tend to have greater gender equality in internet use. In the online only model as well as the online-offline model, the Facebook GG measure is strongly positively correlated with the internet gender gap indicating that in countries where the Facebook use among men and women is more balanced, overall internet use between men and women also tends to be more equal.

Comparing the different regression models, the single variable model with just the Facebook gender gap does better in terms of adjusted R-squared than the offline model which includes a combination of four variables. Furthermore, the use of four offline variables such as GGGR indicators reduces the number of potential countries in our data set for which predictions can be made to 132 countries; contrast this with 152 countries for which predictions can be made when using the online model.

The online-offline model indicates that the adjusted R-squared can be improved if we combine the online Facebook gender gap measure with additional offline variables such as internet penetration and offline GGGR gender gap indicators. However, this model limits the number of countries for which predictions can be made as well as the frequency of updating the predictions, both of which would be important concerns if using the model for nowcasting.

The three models can also be compared on the basis of how well they would be expected to predict for other countries for which the ITU internet gender gap measure is not available using a metric for out-of-sample prediction error, the SMAPE, which was computed using Leave-One-Out cross validation. Table 3 reports the SMAPE for all three models. As can be seen the online-offline model has the lowest error followed by the online model. The offline model has the largest error.

Apart from the prediction accuracy and the number of additional countries for which we can make a prediction, it is also worth looking at the development status of these countries. Out of 48 countries in the UN's "least developed" category, our online model makes predictions for an additional 31 countries, in addition to the three countries for which ITU data are already available. Similarly, using the World Bank's classification of low income countries, our online model makes predictions for an additional

Table 3
Summary of results for three regression models predicting ITU internet Gender Gap Index using (i) a single online Facebook variable; (ii) online and offline variables; (iii) offline variables. Bootstrap estimates of the coefficient standard errors are reported in parentheses.

	Online Model	Onl.-Offl. Model	Offline Model
Intercept	0.933***	0.932***	0.933***
	(0.006)	(0.005)	(0.007)
FB GG (age 18+)	0.071***	0.093***	
	(0.011)	(0.017)	
log(GDP per capita)		0.018*	
		(0.008)	
GGGR – Literacy		−0.018	
		(0.016)	
GGGR – Education		−0.019	
		(0.019)	
Internet Penetration			0.040***
			(0.009)
GGGR – Tertiary Educ.			0.032
			(0.021)
GGGR – Economy			0.043**
			(0.014)
GGGR Score			−0.024
			(0.012)
Adjusted R-squared	0.691	0.791	0.615
Mean Abs. Error	0.0325	0.0288	0.037
SMAPE	3.92%	3.90%	4.97%
F-statistics	169	67.38	29.79
df	74	66	68
N	76	71	73
# predicted countries[a]	152	127	132

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

[a] The number of predicted countries is the number of countries in the dataset that have data on all model variables (so predictions can be made) plus the additional requirement of having more than 200,000 Facebook users in the case of the online and online-offline model which are reliant on the Facebook derived measures of gender gap. If no restrictions are made on the number of Facebook users, predictions can be made for 178 and 134 countries with the online and online-offline models respectively.

22 countries in addition to the 2 countries for which the ITU already reports data. Tables 10 and 11 in the Appendix provide a full breakdown by income/development status and prediction coverage by our online model.

In order to test for the robustness of these fitted online-offline and offline models and their out-of-sample predictive power we compared them to the five model variants fitted using a fivefold cross validation approach described earlier. While the exact choice of offline variables varied between these models they generally either included the same variables as the models in Table 3 or other variables that were strongly correlated with these variables instead. Notably, there was less variability in the online (Facebook) variable selected across the five model variants. This variability in the exact model specification was expected given the strong correlations between several offline variables in the dataset, which is due to the fact that several of these variables (e.g the HDI and GDP per capita, or the various education sub-indices of the GGGR) are measuring very similar things and in some cases can be substituted for the other. However, the out-of-sample prediction of the models in Table 3 was similar to that of the five different model variants. These results indicate that while we must be cautious about the exact model specification and interpretations of coefficients in models using the offline indicators, the predictive performance of the models is not compromised.

Fig. 1 shows the gender gaps in internet access on a world map. Fig. 1a is the Internet Gender Gap Index data computed using the ITU data, while Fig. 1b shows the predictions made by the online Model reported in Table 3. Predictions were made for countries with a Facebook user population of at least 200,000. No prediction was made for China due to a low Facebook penetration of 0.2%, most likely due to internet regulations of the country.[12] (accessed August 14, 2017).

Looking at the maps, for countries where ITU ground truth data are available our prediction generally closely matches it. A noticeable difference between the maps is that with our online model we are able to predict internet gender gaps for a large number of countries in Africa for which no gender gap data from ITU are available. Table 8 in the Appendix provides predicted internet gender gap indicators from different models for all countries in the dataset.

5.2.1. Subsets of countries

The results of the previous sections indicate that all three models explain a significant portion of the variation in the ITU Internet Gender Gap Index as indicated by their high adjusted R-squared values. As R-squared is a quadratic measure, a large value indicates that our models do well in differentiating between "very unequal" and "gender parity" countries, but R-squared is not sensitive to smaller variations.

To evaluate how well our model does at differentiating between countries with high values of the Internet GGI indicating greater gender parity in internet access, we subset the evaluation to countries with a ITU Internet GGI value greater than 0.8 and greater than 0.9. Table 4 shows the results for these countries close to gender parity.

The values in the table were calculated as follows. The fitted regression model was used to make predictions for the subsets of countries where the ground truth ITU Internet GGI was in the specified range. The adjusted R-squared of the regression on the subset of countries was computed as follows:

$$\text{adjusted R-squared} = 1 - (1 - R^2)\frac{n-1}{n-p-1}, \tag{3}$$

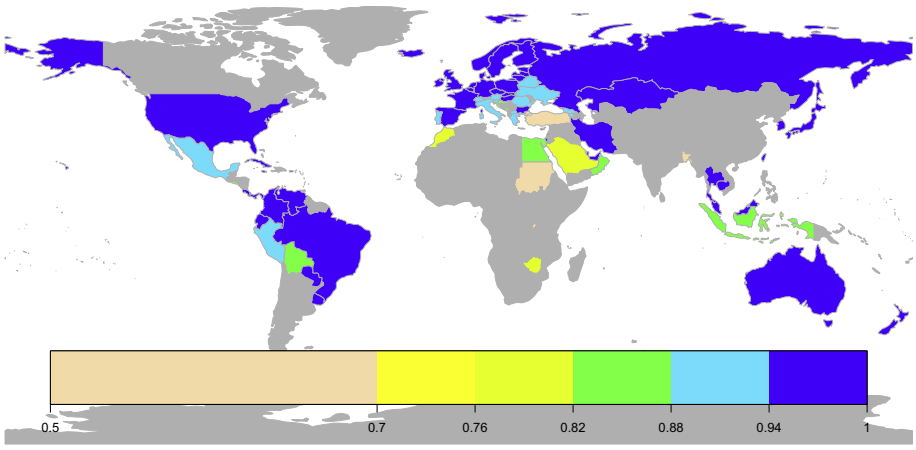
where $n =$ number of countries for which predictions are made, R^2 = square of the Pearson correlation coefficient between predicted and ground truth internet gender gap index and p = the number of model variables in the model (we used $p = 1$ for both models).

As Table 4 shows, the regression model with the Facebook gender gap as predictor as well as the offline regression model with its four variables do very well in distinguishing highly gender equal countries in internet access from more unequal countries but do less well in distinguishing among countries that are close to gender equality in internet access.

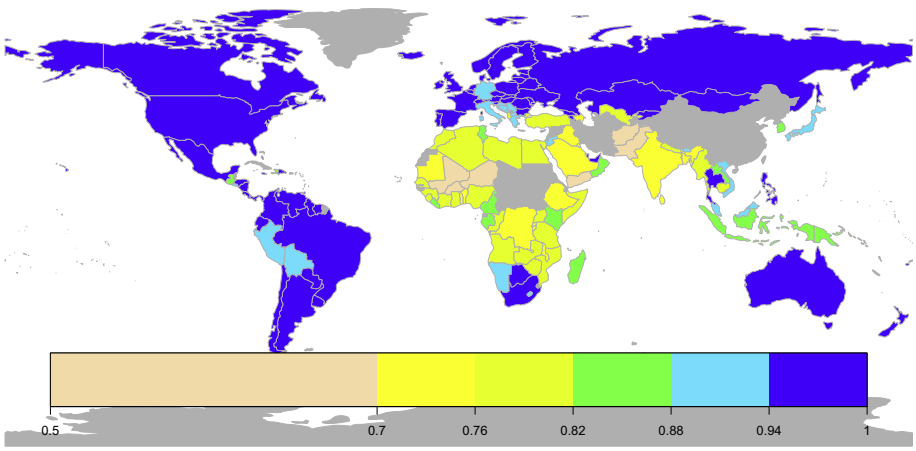
5.2.2. Improving model prediction with correction factor

If every internet user were also a Facebook user, and if Facebook's data were perfectly clean, without any misreported gender or fake accounts, then we would not need to fit models. Rather, we could just read off internet gender gap measures from the Facebook gender gap measures directly. Similarly, if Facebook users were perfectly representative of the wider internet population of the country they are in then, again, the Facebook gender gap would perfectly match the internet gender gap. The reason that there is a strong correlation but not a perfect match between the two measures is mostly due to *selection bias*: in some countries, for example Mexico, women are more likely to be on Facebook than one would expect from their overall internet access, and in other countries such as Saudi Arabia women are less likely to be on Facebook compared to their overall internet access. In fact, most countries fall into the second category, i.e. women are less equal in terms of Facebook usage compared to internet access. If we could understand this selection bias, which is due to a gender difference in likelihood to be a Facebook early or late adopter, we could

[12] http://www.washingtonpost.com/wp-dyn/content/article/2009/07/07/AR2009070701162.html



(a) Internet Gender Gap Index computed using ITU ground truth data



(b) Internet Gender Gap Index predicted from our Online Model using the Facebook 18+ user Gender Gap Index

Fig. 1. The internet gender gap index computed using (a) ITU ground truth data, and (b) predicted using Facebook data.

Table 4
Adjusted R-squared of the Online and Offline models for predictions made on subsets of countries in the dataset. Table also indicates number of countries on which this prediction was made.

	Online model	Offline model
Internet GGI > 0.8	adjusted R-squared 0.433 predicted on 71 countries	adjusted R-squared 0.317 predicted on 68 countries
Internet GGI > 0.9	adjusted R-squared 0.0733 predicted on 60 countries	adjusted R-squared 0.0790 predicted on 57 countries

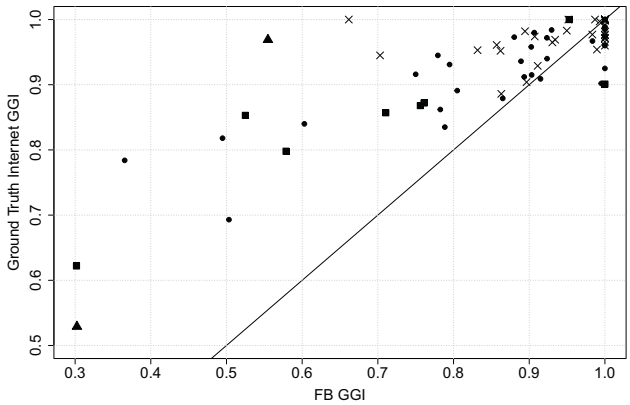
"de-bias" our data, leading to an improved prediction accuracy. A similar approach was explored in Yazdani and Manovich (2012), correcting for the fact that early email adopters are more likely to be international migrants.

Fig. 2 is a scatter plot of the Facebook gender gap index for ages 18+ against the Internet GGI with a diagonal line. The points are coded based on the quartile of the distribution of the internet penetration (Fig. 2a) or the Literacy GGGR (Fig. 2b) they fall under. As can be seen, the closed Facebook gender gap is often lower than the ITU ground truth, i.e. Facebook data indicates a larger gender disparity than the actual disparity in internet access. However, the extent of this bias differs for different countries. From the Figure it is evident that in countries with lower levels of internet penetration or Literacy GGGR (indicating large disparities in literacy between male and female) the extent of the bias is generally larger.

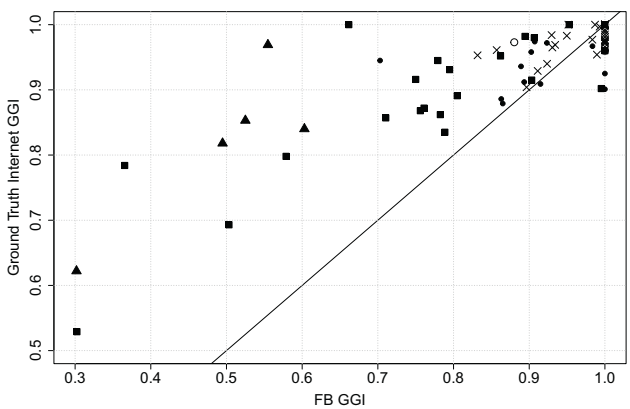
This suggests that the bias in the Facebook derived measure of the internet gender gap can be corrected by multiplication with a correction factor that is a function of some other variable such as the internet penetration. This bias corrected Facebook gender gap can then be used in fitting a regression model to predict the internet gender gap. A possible functional form for the correction factor is:

$$\text{Correction Factor} = 1 + k * (1 - x) \tag{4}$$

In Eq. (4), x is the variable being used for bias correction, which in this case is assumed to be a variable in the interval $[0, 1]$, for example internet penetration. Once this value reaches 1.0 we assume that the need for a bias correction factor disappears. k is a tunable parameter whose sign and magnitude indicate whether and by how much the Facebook gender gap is to be de-biased by decreasing or increasing its value.



(a) FB GGI against Internet GGI with points coded according to the Internet Penetration



(b) FB GGI against Internet GGI with points coded according to the GGGR for Literacy

Fig. 2. Internet GGI versus the Facebook GGI. Points are labeled according to the quartile of the distribution of Internet penetration (top panel) or the the Literacy GGGR (bottom panel) they fall in. Points falling in the first through the fourth quartiles are labeled as triangle, square, circle and cross respectively. Empty circle indicates missing value for that variable.

The variables x above using internet penetration is an example of one such factor. Other variables, such as gender gaps in literacy, that are intuitively or theoretically related to the early adopter gender bias on Facebook, can also be used. In countries where the gender gap in literacy is far from being closed, we would expect that internet access itself might not be enough to enable women to participate fully online. As a result, women's visibility on platforms such as Facebook, which requires a certain minimum level of verbal and digital literacy, would be lower than those where the literacy gender gap has closed.

We experimentally tried different choices of variables x. For each choice of x, values of the tunable parameter in the range $[-10, 10]$ at intervals of 0.1 were tried, each time correcting the Facebook gender gap measure for that choice of x and k and estimating a regression model to predict the internet gender gap index as a function of the corrected Facebook gender gap. The best choice of the tunable parameter was the one which resulted in the largest adjusted R-squared. Overall, the literacy gender gap works best as the correction variable resulting in an increase in the adjusted R-squared of the single variable online model from 0.691 (uncorrected) to 0.730 (corrected).

However, this general approach raises two concerns. The first relates to a potential sensitivity of the choice of parameters, i.e. a risk of overfitting. In the case of using the literacy gender gap, the adjusted R-squared of the model was quite robust to small changes in the value of the tuning parameter k around its optimal value.[13] The second concern is the loss in coverage in terms of the number of countries for which predictions can be made as we require additional data, in this case literacy gender gap scores, in addition to the more widely available Facebook data to be able to make predictions.

5.3. Regression models predicting mobile phone gender gaps

The previous sections focused on using Facebook data to derive measures of the internet gender gap. Given the wide range of possible ad targeting options on Facebook, which include targeting by device type, one could imagine using data on the number of male and female mobile device users to derive measures of the mobile phone gender gap using the same procedures which were used for the internet gender gap. However, the scarcity of available ground truth data – mobile ownership data by gender from the GSMA were only available for 22 countries – limits the scope of the conclusions that can be made. This section presents the results of several regression models for predicting the mobile phone gender gap.

Table 5 shows the results of the three different regression models with standardized coefficients. The online model is a single variable model using the most strongly correlated Facebook gender gap measure, in this case the Facebook gender gap for people aged 25–29, as the predictor. The offline model replicates that reported by GSMA and consists of the three variables, namely the Oxford Multidimensional Poverty Index, an indicator variable for being in South Asia[14] and an indicator for a low unique subscriber penetration defined in Section 3.3. Our analysis here evaluates whether data from Facebook, or the use of Facebook in addition with offline indicators can improve the predictive power of the offline model from the GSMA report.

The online-offline model was arrived at using the greedy stepwise forward method which was used earlier with the extra condition that for the added variable there should be at least twenty countries that have data for that variable and the ground truth in order to safeguard against fitting models on too small amounts of data. The variables considered in building this model were the variables in offline model, with the exception of the indicator variable for being in South Asia, and the online Facebook derived gender gap measures for ages 25–29 and for various device types. The justification for excluding the indicator for South Asia were that (i) it is a static variable that does not change over time, reducing its value for tracking temporal changes, and (ii) it is particular to a single geographic region, indicating a risk for overfitting as no potential causal explanation for the particular choice is given in the original GSMA report.

As can be seen in Table 5, the offline model achieves the highest adjusted R-squared followed closely by the online-offline model. The online model, using just a single Facebook derived measure of gender gap still achieves an adjusted R-squared greater than 0.5. The SMAPE error in the Table reports an estimate of out-of-sample error. As can be seen the error rates are generally larger than for the model predicting the internet GGI which is not surprising given the small size of the datasets. In this case the offline model has the lowest error followed by the online-offline model

[13] When using the gender gap in literacy as the variable used for correction, the optimal value of the tuning parameter was $k = 2.0$. However, using values of k in the range $[1.0, 3.0]$ the adjusted R-squared of the regression model with the corrected Facebook gender gap ranged between 0.715 and 0.730.

[14] This indicator takes value 1.0 if the country is located in South Asia as per World Bank's classification of world regions and a value of 0.0 otherwise.

Table 5
Summary of results for three regression models predicting GSMA Mobile Phone Gender Gap Index using (i) a single online Facebook variable; (ii) online and offline variables; (iii) offline variables. Bootstrap estimates of the coefficient standard errors are reported in parentheses. Note: USP stands for Unique Subscriber Penetration, OMPI stands for Oxford Multi-dimensional Poverty Index.

	Online Model	Onl.-Offl. Model	PCA	Offline Model
Intercept	0.820***	0.852***	0.820***	0.893***
	(0.024)	(0.022)	(0.024)	(0.014)
FB GGI	0.116**	0.160***		
(age 25–29)	(0.029)	(0.045)		
Low USP		−0.166		−0.200
Indicator		(0.052)		(0.043)
FB iOS devices		−0.084		
users GGI		(0.038)		
OMPI				−0.046
				(0.018)
South Asia				−0.241
Indicator				(0.035)
First PC			0.084***	
			(0.016)	
Adj. R-squared	0.524	0.798	0.584	0.874
Mean Abs. Error	0.091	0.056	0.086	0.040
SMAPE	13.10%	8.64%	11.91%	6.72%
F-statistics	24.12	27.29	29.05	47.17
df	20	17	19	17
N	22	21	21	21
# pred. countries[a]	153	62	62	69

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.
[a] The number of predicted countries is the number of countries in the data set that have data on all model variables (so predictions can be made) plus the additional requirement of having more than 200,000 Facebook users in the case of the Online and Online-Offline model. If no restrictions are made on the number of Facebook users, predictions can be made for 179 and 99 countries with the Online and Online-Offline models respectively.

and finally the online model which has the largest error. Similar to the internet gender gap models, Table 5 highlights the trade-offs in the choice of model such as the ability to make predictions for a large sample of countries using the online model at the sacrifice of prediction accuracy as demonstrated by the adjusted R-squared and SMAPE. Table 8 in the Appendix provides predicted mobile phone gender gap indicators from different models for all countries in the dataset.

Given the small number of data points, we must be cautious of potential over-fitting as the resulting model may not generalize well when predicting for countries where these data are not available. This risk, while still present, is smaller for simpler models. While the online model uses one variable, the online-offline and offline models use multiple variables for the prediction task. We estimate a parsimonious version of the online-offline model using Principal Component Analysis (PCA). Using PCA, we reduce the number of variables in the model by fitting on a linear combination of the three variables (a principal component) from the online-offline model reported in Table 5 which also reports the model resulting from this procedure.

The one variable online-offline model that uses the first principal component attains a larger adjusted R-squared (0.584 versus 0.524) and a smaller SMAPE (11.91% versus 13.10%) than the one for the online model reported in Table 5. Although the risk of over-fitting is lower, the PCA model for the online-offline model shows poorer within- and out-of-sample model performance compared with the three variable online-offline model.

As with the internet gender gap models, the online-offline mobile gender gap model in Table 5 was compared to the five variants obtained using a cross validation approach discussed before. Once again, while the exact model variables varied from model to model, all models included the indicator variable for a low USP and most included some Facebook gender gap index. However, the out-of-sample prediction error of the online-offine model was comparable to that of these model variants. This suggests while the exact model specifications may vary, the models presented in Table 5 provide relatively good prediction performance.

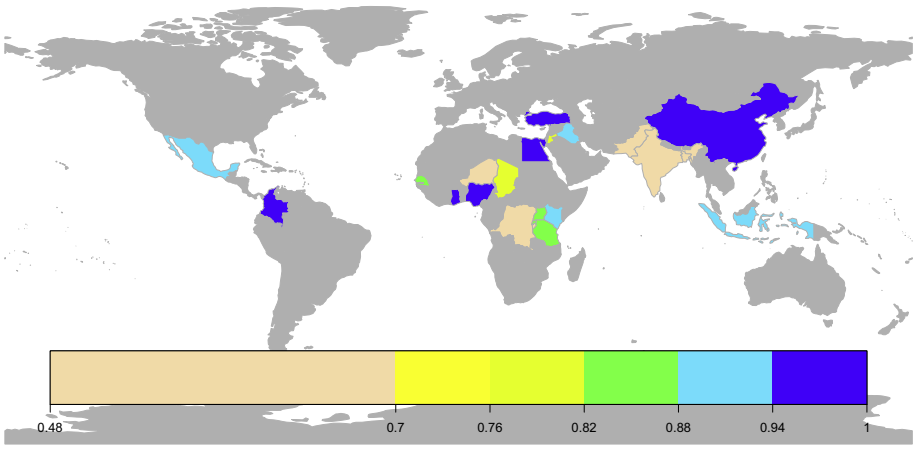
Fig. 3 shows the mobile phone GGI from GSMA data in panel 3a and the predictions made by the Online model in panel 3b. The criteria for making predictions were the same as for the maps on the internet GGI, namely a Facebook user base of at least 200,000, with no prediction made for China. As seen on the maps, the use of the online model increases the geographic coverage of countries, filling in gaps in Africa.

6. Trends over time

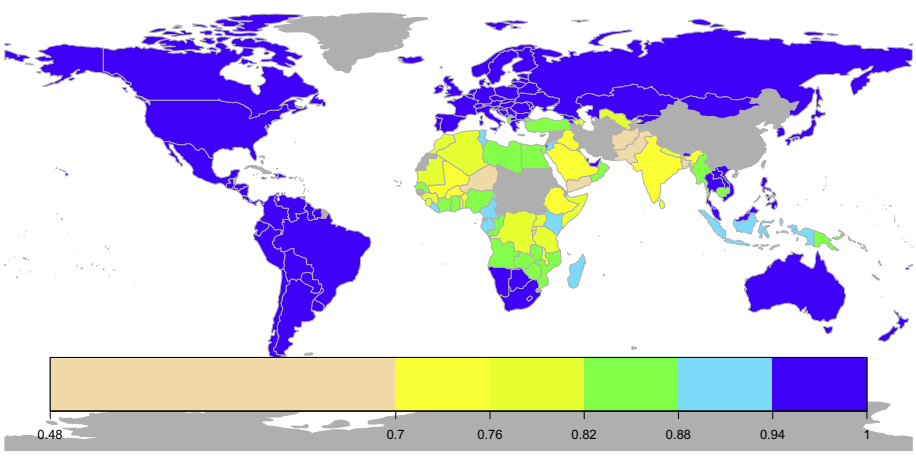
The previous sections explored the performance of different models in predicting the Internet and Mobile Phone GGI using a combination of online and offline variables. The higher temporal resolution of online (Facebook) variables make them particularly useful for nowcasting. These data can be regularly collected over short time intervals and allow us to measure changes over shorter time scales than would be possible with other data sources. In this section we present a first examination of the temporal variation over a few months in the size of the user estimates from the Facebook data source and in the predictions made by the online models.

6.1. Temporal trends in audience size by gender

Table 6 reports a summary of the percentage change in the size of the female and male user numbers (or audience size) over different time periods for the countries in the dataset. There is some variability in the estimated audience size for the different countries. Whereas median change over the period from June to September 2017 was 0, median change between September and October was 5.88%. This may be due to the change in user numbers corresponding to the start of the academic year. In general, however, percentage changes in audience size in most countries were fairly small and when they did occur, occurred for both male and female users.



(a) Mobile Phone Gender Gap Index computed using GSMA ground truth data



(b) Mobile Phone Gender Gap Index predicted from Online model using the Facebook age 25-29 user Gender Gap Index

Fig. 3. The Mobile Phone Gender Gap Index computed using (a) GSMA Ground Truth data, and (b) predicted using Facebook data.

Table 6
Summary statistics of the percentage change in the estimated audience size for a period of several months.

		Percentage Change in Audience Size		
Targeted Audience	Period	10th Percentile	Median	90th Percentile
Female age 18+	Jun-Sep	−6.71	0.00	4.79
Female age 18+	Sep-Oct	0.00	5.88	14.29
Female age 18+	Oct-Nov	0.00	1.64	7.02
Male age 18+	Jun-Sep	−9.11	−1.34	2.74
Male age 18+	Sep-Oct	0.00	5.97	13.33
Male age 18+	Oct-Nov	0.00	0.00	5.88
Female/Male Ratio	Jun-Sep	−2.87	1.16	7.83
Female/Male Ratio	Sep-Oct	−4.35	0.00	5.31
Female/Male Ratio	Oct-Nov	−3.13	0.00	5.00

6.2. Temporal trends in online model predictions

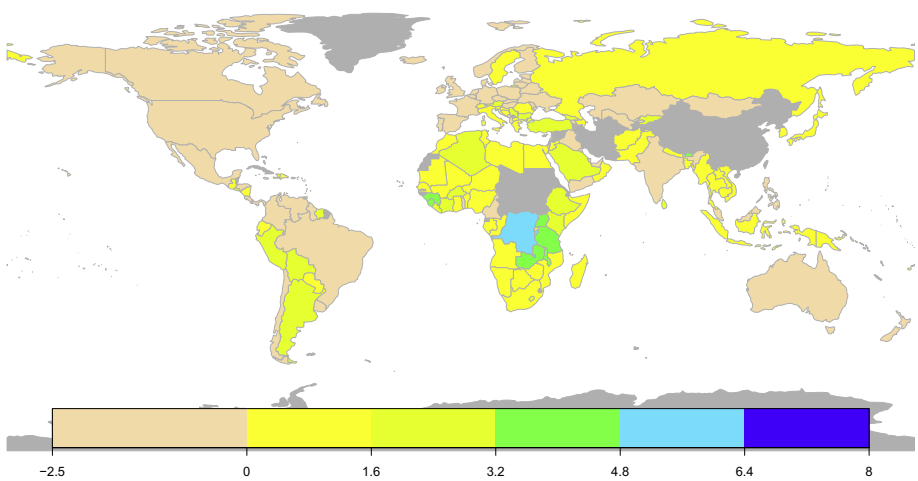
Table 7 summarizes the percentage change in the predicted Internet and Mobile GGI using the online model over various time periods. The changes shown are relative changes over the time period, e.g for change between June and September this would be $\frac{Predicted\ GGI\ in\ Sep - Predicted\ GGI\ in\ Jun}{Predicted\ GGI\ in\ Jun}$. The same ground truth ITU Internet and GSMA Mobile Phone Gender Gap Indices used as outcome variables in the models presented in Tables 3 and 5 are used for these new models. Unlike the actual audience sizes, the predicted indices

experienced much smaller percentage changes. Temporal changes in the audience size and the model predictions could be due to changes in the underlying variable we are trying to measure (the online presence of males versus females) and also due to random fluctuations (e.g start of school years) or noise due to such things as measurement errors. We would not expect drastic changes in a country's digital gender gaps to occur over a short time period, and it is likely that random fluctuations will dominate any observed changes. These results indicate that the model predictions are relatively stable to these random fluctuations as the predicted index value change by very small percentages.

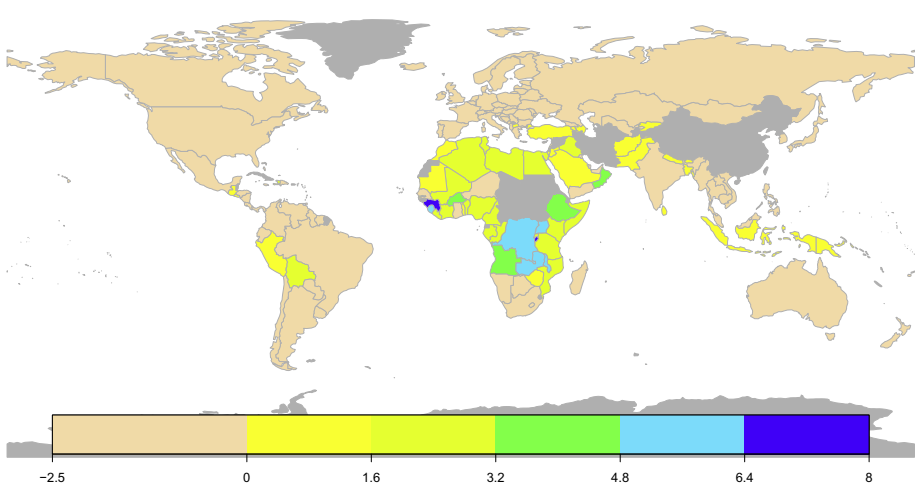
Fig. 4 shows a map with the relative percentage change between June and November 2017 in the predicted Internet GGI and Mobile Phone GGI values from the online model using Facebook data. In most countries, the map highlights that percent changes in the predicted values were relatively small. Larger,

Table 7
Summary statistics of the percentage change in the predicted Internet and Mobile Phone GGI using the online model for a period of several months.

		Percentage Change in Prediction		
Variable	Period	10th Percentile.	Median	90th Percentile
Internet GGI	Jun-Sep	−0.365	0.008	2.064
Internet GGI	Sep-Oct	−1.037	0.000	1.445
Internet GGI	Oct-Nov	−0.770	0.000	1.305
Mobile Phone GGI	Jun-Sep	−0.006	0.000	2.627
Mobile Phone GGI	Sep-Oct	−0.903	0.000	0.623
Mobile Phone GGI	Oct-Nov	−0.795	0.000	1.011



(a) Percentage Change in the Internet GGI Jun-Nov 2017



(b) Percentage Change in the Mobile GGI Jun-Nov 2017

Fig. 4. Percentage Change in the Internet GGI (top panel) and Mobile Phone GGI (bottom panel) predicted by the Online models over the period from June to November 2017.

positive changes in the predicted GGIs suggestive of progress in closing gender gaps were observed in sub-Saharan Africa. While this seems encouraging, regular data collection over a longer period such as a year is necessary to allow us to rule out periodic patterns, such as start of academic years, when measuring changes in the GGI values with greater certainty.

7. Discussion and conclusion

We began this paper by identifying a significant data gap in an important development indicator – gender gaps in internet and mobile access. Although the role of ICT as a tool for women's empowerment, and issues of gender equity in internet and mobile access have been widely recognized among development organizations and scholars, existing infrastructure to monitor gender gaps in ICT use particularly in low-income countries is limited. In this setting, we evaluated the potential contribution of a source of digital trace 'big data', specifically data on user estimates provided by Facebook's advertising platform, for monitoring and nowcasting digital gender gaps.

The results were very promising with a regression model based solely on Facebook's gender gap explaining 69% of the variance in the ground truth Internet Gender Gap Index that we computed using ITU data.

In the absence of Facebook data, these results could not be matched even when combining several offline measures of a country's development and levels of gender inequality. Although models with the Facebook measure alone performed well, models with the most predictive power combined both online and offline variables, either as multivariate models or through the use of bias-correction factors. Through our approach, we have sought to illustrate different ways in which a unique source of digital trace data, which although purposed for advertising, can also valuably be used for monitoring digital gender gaps.

Each of the different data sources that we have used come with their strengths and weaknesses, which ultimately, determine the choice of the appropriate model. The choice of variables limits the number of countries for which we are able to generate predictions. For example, the Global Gender Gap Report indicators are available for approximately 140 countries and are updated annually. So although the model has better predictive power, a model with literacy gender gap correction could make predictions for fewer countries than a single variable model using just the Facebook gender gap. Moreover, using just the Facebook data, predictions could be made on a monthly or even daily basis, as we tried to show in Section 6. In addition to their frequently updating nature, the Facebook data further allow for disaggregation by other socio-demographic characteristics of the user (e.g. education) or by subnational geographies, which is a level of granularity not permitted by ITU or GSMA data on digital gender gaps. We have chosen not to present results with those attributes in this paper as limited availability of other data sources on those fronts to enable us to validate our results in the same way that we are able to do for the country-level analysis would require systematic evaluation of a number of additional data sources. This is promising area for further research.

Note, however, that our online model still relies on access to data for the population gender imbalance. Though in most countries the ratio of the number of women to number of men is close to 1.0, there are exceptions ranging from 1.25 in the case of Martinique to 0.29 in the case of United Arab Emirates for the 18 + population (United Nations Population Division, 2017).

Furthermore, though the actual population size might change rapidly, the gender ratio typically changes on much slower time scales. Hence we do not see the reliance on the population gender distribution as an important limitation.

It is nevertheless important to acknowledge the drawbacks of the Facebook data source. Since the Ad Audience Estimates are targeted to advertisers, their documentation does not clearly specify the algorithm that determines the estimates of the users. As these data are not primarily targeted for research, it is also possible that the algorithm may change without notice. It must be emphasized that 'big data' sources, such as the ones we have used here, cannot be a replacement for data collected from statistically robust probability samples in household surveys and censuses, such as those collected by national and international statistical agencies but instead serve as a complement to them. Another shortcoming of our approach is that it is better at tracking relative measures – in this case, female-to-male ratios rather than absolute levels of internet adoption by gender, which is what official statistics such as those provided by the ITU are able to provide.

Whereas the 2016 ITU report that we use for validating our Facebook measures provides data from 2011–15, the Facebook data that we used in our analyses are more recent. Differences between the predictions and the ground truth could potentially be explained by changes over time. As Facebook's advertising platform does not provide historic data, we will only know for sure in due course as we continue to collect data to allow for tracking of change over time. The relationship between offline gender gap and other development indicators, as well as the Facebook data and the ITU data could plausibly change over time. In this case, generating predictions and nowcasting digital gender gaps using Facebook data will require re-calibrated models as new data become available. Given the short timelines of the sustainable development goals, our approach using the Facebook data nevertheless provides a method to benchmark and generate more frequent estimates to help with regular monitoring of progress related to gender equality.

Conflict of interest

None.

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). We would like to thank Rebecca Furst-Nichols and Bapu Vaitla for pointing us to the Data2X Big Data for Gender Challenge, which helped us to better understand the need for this type of research. The paper started when Masoomali Fatehkia was a visiting researcher at Qatar Computing Research Institute with funding from the Princeton University International Internships Program and the Center for Information Technology Policy Norman B. Tomlinson Jr., '48 Scholarship. We are grateful for the helpful comments of three anonymous reviewers and the editor of World Devlopment.

Appendix A. Detailed results

Table 8
The internet and mobile phone GGI, where available, for each country in the data set alongside the GGI predicted by each of the six models. *Note:* Model predictions were truncated at 1.0 in keeping with the GGGR methodology.

	Country	Ground Truth Internet GG	Internet online model prediction	Internet Online-Offline model prediction	Internet Offline model prediction	Gound Truth Mobile GG	Mobile Online model prediction	Mobile Online-Offline model prediction	Mobile Offline model prediction
1	Afghanistan	–	.654	–	–	–	.623	–	–
2	Albania	–	.799	.741	.929	–	.870	–	–
3	Algeria	–	.802	.828	.827	–	.807	.851	.945
4	Andorra	–	–	–	–	–	–	–	–
5	Angola	–	.813	1.00	.764	–	.834	.717	–
6	Argentina	–	.957	.942	.930	–	1.00	1.00	–
7	Armenia	.936	.941	.901	.961	–	.953	–	–
8	Aruba	–	.971	–	–	–	1.00	–	–
9	Australia	1.00	.986	.993	1.00	–	1.00	.979	–
10	Austria	.904	.944	.952	.959	–	1.00	–	–
11	Azerbaijan	–	.753	.681	.988	–	.770	–	–
12	Bahrain	1.00	.849	–	.947	–	.920	.813	–
13	Bangladesh	.622	.702	.658	.694	.620	.693	.734	.653
14	Barbados	–	.986	.972	1.00	–	1.00	–	–
15	Belarus	.925	.986	.974	.974	–	1.00	–	–
16	Belgium	.983	.966	.966	.989	–	1.00	–	–
17	Belize	–	.986	.968	.916	–	1.00	–	–
18	Benin	–	.741	1.00	.691	–	.792	.882	.864
19	Bhutan	–	.914	1.00	.795	–	.970	–	–
20	Bolivia	.868	.887	.874	.793	–	.983	.967	.922
21	Bosnia and Herzegovina	–	.931	.906	.916	–	1.00	–	–
22	Botswana	–	.949	.923	.917	–	1.00	1.00	–
23	Brazil	.985	.986	.972	.935	–	1.00	1.00	.941
24	Brunei	–	.910	.921	.974	–	.955	–	–
25	Bulgaria	.972	.955	.944	.932	–	1.00	–	–
26	Burkina Faso	–	.698	.896	.698	–	.736	.819	.803
27	Burundi	.529	.702	.643	.666	.580	.691	.526	.624
28	Cambodia	.969	.805	.864	.739	–	.869	.882	.907
29	Cameroon	–	.847	.950	.784	–	.915	.996	.880
30	Canada	–	.986	.992	–	–	1.00	–	–
31	Cape Verde	–	.934	.934	.857	–	1.00	1.00	–
32	Central African Republic	–	.727	–	–	–	.752	.692	.631
33	Chad	–	.655	.996	.584	.810	.661	–	–
34	Chile	–	.981	.975	.926	–	1.00	.992	–
35	China	–	.833	.818	.916	.990	.941	.937	.942
36	Colombia	.974	.986	.970	.944	.970	1.00	1.00	.940
37	Comoros	–	.824	–	–	–	.892	.833	.700
38	Costa Rica	.992	.986	.972	.941	–	1.00	1.00	–
39	Cote d'Ivoire	–	.767	1.00	.714	–	.823	.898	.863
40	Croatia	.879	.931	.915	.957	–	1.00	–	–
41	Curacao	–	–	–	–	–	–	–	–
42	Cyprus	.958	.947	.939	.954	–	.993	–	–
43	Czech Republic	.971	.986	.987	.980	–	1.00	–	–
44	Dem. Republic of the Congo	–	.733	–	–	.670	.766	.668	.639
45	Denmark	1.00	.986	.994	1.00	–	1.00	–	–
46	Djibouti	–	.787	–	–	–	.846	.782	.709
47	Dominica	–	–	–	–	–	–	–	–
48	Dominican Republic	–	.974	.965	.926	–	1.00	1.00	.937
49	Ecuador	.982	.943	.920	.949	–	1.00	1.00	.943
50	Egypt	.853	.793	.834	.789	.980	.843	.862	.942
51	El Salvador	.872	.889	.861	.887	–	.982	–	–
52	Equatorial Guinea	–	.752	–	–	–	.767	.842	–
53	Eritrea	–	.859	–	–	–	.998	.960	–
54	Estonia	.977	.986	.988	.984	–	1.00	–	–
55	Ethiopia	–	.719	.845	.638	–	.745	.604	.595
56	Fiji	–	.973	–	–	–	1.00	–	–
57	Finland	1.00	.986	.991	.980	–	1.00	–	–
58	France	.977	.980	.982	.959	–	1.00	–	–
59	Gabon	–	.840	–	–	–	.909	.975	.927
60	Georgia	.912	.943	.916	.940	–	1.00	–	–
61	Germany	.953	.918	.936	.945	–	1.00	–	–
62	Ghana	–	.786	.802	.800	.960	.870	.985	.904

(*continued on next page*)

Table 8 (*continued*)

	Country	Ground Truth Internet GG	Internet online model prediction	Internet Online-Offline model prediction	Internet Offline model prediction	Gound Truth Mobile GG	Mobile Online model prediction	Mobile Online-Offline model prediction	Mobile Offline model prediction
63	Greece	.931	.903	.899	.937	–	.992	–	–
64	Grenada	–	.986	–	–	–	1.00	–	–
65	Guatemala	–	.870	.896	.866	–	.971	–	–
66	Guinea	–	.772	1.00	.686	–	.813	.892	.823
67	Guinea–Bissau	–	.735	–	–	–	.775	.718	.647
68	Guyana	–	.986	–	–	–	1.00	–	–
69	Haiti	–	.862	–	–	–	.955	1.00	.880
70	Honduras	–	.986	.947	.851	–	1.00	1.00	.927
71	Hong Kong	.960	.986	–	–	–	1.00	1.00	–
72	Hungary	.954	.982	.984	.983	–	1.00	–	–
73	Iceland	.997	.983	.991	.990	–	1.00	–	–
74	India	–	.709	.730	.754	.640	.721	.746	.654
75	Indonesia	.857	.868	.837	.885	.900	.901	.840	.929
76	Iraq	–	.721	–	–	.890	.732	.805	.934
77	Ireland	1.00	.986	1.00	.970	–	1.00	–	–
78	Israel	.973	.938	–	.955	–	.986	.933	–
79	Italy	.886	.931	.923	.935	–	1.00	–	–
80	Jamaica	1.00	.986	.960	.927	–	1.00	–	–
81	Japan	.961	.928	.924	.944	–	1.00	1.00	–
82	Jordan	–	.882	.834	.878	.790	.911	.865	.945
83	Kazakhstan	.974	.986	.985	.976	–	1.00	–	–
84	Kenya	–	.833	.824	.886	.930	.899	.882	.896
85	Kiribati	–	.986	–	–	–	1.00	–	–
86	Kosovo	–	–	–	–	–	–	–	–
87	Kuwait	–	.814	–	.936	–	.830	.839	–
88	Kyrgyzstan	–	.945	.897	.871	–	.992	–	–
89	Laos	–	.866	.916	.876	–	.962	–	–
90	Latvia	.976	.986	.981	1.00	–	1.00	–	–
91	Lebanon	–	.856	.854	.910	–	.880	.785	–
92	Lesotho	–	.937	.874	.843	–	1.00	1.00	.909
93	Liberia	–	.832	1.00	.693	–	.911	.831	.646
94	Libya	–	.813	–	–	–	.824	.843	.945
95	Liechtenstein	–	–	–	–	–	–	–	–
96	Lithuania	1.00	.981	.977	.993	–	1.00	–	–
97	Luxembourg	.969	.960	.974	1.00	–	1.00	–	–
98	Macau	.967	–	–	–	–	1.00	–	–
99	Macedonia	.916	.885	.863	.942	–	.960	–	–
100	Madagascar	–	.844	.798	.818	–	.909	.711	.650
101	Malawi	–	.735	.750	.750	–	.744	.565	.679
102	Malaysia	.945	.897	.893	–	–	.952	.864	–
103	Maldives	–	.772	.703	.940	–	.791	–	–
104	Mali	–	.690	1.00	.649	–	.701	.787	.824
105	Malta	.965	.958	.996	.952	–	1.00	–	–
106	Marshall Islands	–	–	–	–	–	–	–	–
107	Mauritania	–	.754	.911	.621	–	.776	.724	.870
108	Mauritius	.835	.900	.893	.904	–	.990	.983	–
109	Mexico	.902	.984	.984	.868	.940	1.00	.995	.945
110	Micronesia	–	.986	–	–	–	1.00	–	–
111	Moldova	–	.986	.954	.937	–	1.00	–	–
112	Monaco	–	–	–	–	–	–	–	–
113	Mongolia	–	.986	.973	.940	–	1.00	–	–
114	Montenegro	.862	.898	.873	.927	–	1.00	–	–
115	Morocco	.818	.781	.859	.836	–	.803	.845	.928
116	Mozambique	–	.782	.931	.748	–	.842	.911	.842
117	Myanmar	–	.791	–	–	–	.850	.850	.910
118	Namibia	–	.934	.897	.861	–	1.00	1.00	.895
119	Nauru	–	–	–	–	–	–	–	–
120	Nepal	–	.777	.852	.767	–	.786	.809	.671
121	Netherlands	1.00	.986	1.00	.971	–	1.00	–	–
122	New Zealand	1.00	.986	.990	.994	–	1.00	.928	–
123	Nicaragua	–	.945	.897	–	–	1.00	.999	.927
124	Niger	–	.664	–	–	.550	.655	.546	.584
125	Nigeria	–	.817	1.00	.844	.950	.864	.921	.865
126	Norway	.998	.986	.998	1.00	–	1.00	–	–
127	Oman	.840	.825	–	.909	–	.841	.759	–
128	Pakistan	–	.696	.907	.762	.490	.697	.550	.443
129	Palau	–	–	–	–	–	–	–	–
130	Panama	1.00	.986	.991	.960	–	1.00	.995	–
131	Papua New Guinea	–	.833	–	–	–	.867	–	–
132	Paraguay	1.00	.967	.946	.902	–	1.00	.975	–
133	Peru	.891	.907	.898	.899	–	.966	.898	.935
134	Philippines	–	.986	.957	.917	–	1.00	1.00	.932
135	Poland	.967	.980	.975	.938	–	1.00	–	–

Table 8 (*continued*)

	Country	Ground Truth Internet GG	Internet online model prediction	Internet Online-Offline model prediction	Internet Offline model prediction	Gound Truth Mobile GG	Mobile Online model prediction	Mobile Online-Offline model prediction	Mobile Offline model prediction
136	Portugal	.915	.947	.952	.943	–	1.00	–	–
137	Qatar	.974	.949	.982	1.00	–	1.00	1.00	–
138	Republic of the Congo	–	.775	–	–	–	.823	.931	.898
139	Romania	.909	.952	.944	.935	–	1.00	–	–
140	Russia	.984	.958	.944	.973	–	1.00	–	–
141	Rwanda	–	.731	.671	.809	.820	.750	.791	.877
142	Saint Kitts and Nevis	–	–	–	–	–	–	–	–
143	Saint Lucia	–	.986	–	–	–	1.00	–	–
144	St. Vincent & the Grenadines	–	.986	–	–	–	1.00	–	–
145	Samoa	–	.986	–	–	–	1.00	–	–
146	San Marino	–	–	–	–	–	–	–	–
147	Sao Tome and Principe	–	.861	–	–	–	.946	1.00	.925
148	Saudi Arabia	.784	.728	.717	.833	–	.736	.686	–
149	Senegal	–	.771	.962	.701	.850	.840	.937	.858
150	Serbia	–	.911	.883	.931	–	1.00	–	–
151	Seychelles	–	.920	–	–	–	.967	.946	–
152	Sierra Leone	–	.784	–	–	–	.805	.880	.822
153	Singapore	.952	.930	.971	–	–	.984	.904	–
154	Sint Maarten	–	–	–	–	–	–	–	–
155	Slovakia	.981	.986	.985	.974	–	1.00	–	–
156	Slovenia	.940	.955	.945	.962	–	1.00	–	–
157	Solomon Islands	–	.909	–	–	–	.980	–	–
158	Somalia	–	.802	–	–	–	.803	.701	.608
159	South Africa	–	.963	.952	.884	–	1.00	1.00	.937
160	South Korea	.945	.865	.866	.872	–	1.00	1.00	–
161	Spain	.960	.986	.995	.944	–	1.00	–	–
162	Sri Lanka	–	.739	.663	.823	–	.781	.867	–
163	Suriname	–	.951	.931	–	–	1.00	–	–
164	Swaziland	–	.869	.808	.856	–	.944	.913	.929
165	Sweden	1.00	.967	.969	1.00	–	1.00	–	–
166	Switzerland	.929	.950	.958	.980	–	1.00	–	–
167	Taiwan	.956	–	–	–	–	–	–	–
168	Tajikistan	–	.747	.685	.754	–	.741	–	–
169	Tanzania	–	.789	.801	.646	.870	.819	.877	.870
170	Thailand	.980	.948	.945	.967	–	1.00	1.00	.945
171	The Bahamas	–	.986	.979	–	–	1.00	–	–
172	The Gambia	–	.749	.804	.751	–	.791	.869	.860
173	Timor–Leste	–	.828	–	–	–	.854	–	–
174	Togo	–	.720	–	–	–	.763	.695	.678
175	Tonga	–	.986	–	–	–	1.00	–	–
176	Trinidad and Tobago	–	.986	.995	–	–	1.00	–	–
177	Tunisia	–	.864	–	–	–	.929	.955	.945
178	Turkey	.693	.784	.779	.828	.980	.852	.907	.939
179	Turkmenistan	–	.711	–	–	–	.803	–	–
180	Tuvalu	–	–	–	–	–	–	–	–
181	Uganda	–	.786	.807	.782	.860	.814	.829	.848
182	Ukraine	.901	.986	.959	.913	–	1.00	–	–
183	United Arab Emirates	.972	.986	1.00	.930	–	1.00	1.00	–
184	United Kingdom	.966	.986	.992	.981	–	1.00	–	–
185	United States	1.00	.986	.997	1.00	–	1.00	–	–
186	Uruguay	.989	.986	.978	.948	–	1.00	1.00	–
187	Uzbekistan	–	.809	–	–	–	.805	–	–
188	Vanuatu	–	.944	–	–	–	1.00	–	–
189	Venezuela	1.00	.986	–	.932	–	1.00	1.00	–
190	Vietnam	–	.910	.890	.937	–	1.00	.998	.938
191	Yemen	–	.646	.898	.615	–	.646	.730	.883
192	Zambia	–	.813	–	–	–	.849	.854	.871
193	Zimbabwe	.798	.815	.747	.846	–	.861	.887	.905

Appendix B. Variables in the dataset

Table 9
Full list of online and offline variables compiled in the dataset. Correlations of the different variables with the ground Truth ITU internet gender gap index and ground truth GSMA gender gap index; the number in parenthesis indicates the number of data points used to compute the correlation. The year in parenthesis indicates latest year of data availability. Some variables such as the OMPI are computed based on combined data from multiple sources which may add to the latency of the variable. *Note:* GGI stands for Gender Gap Index, FB stands for Facebook, OMPI stands for Oxford Multidimensional Poverty Index, USP stands for unique subscriber penetration, GGGR refers to data on gender gap measures from the Global Gender Gap Report.

Variable	Number of Countries with data	Correlation with ITU Internet GGI	Correlation with GSMA Mobile Phone GGI	Latency/(Latest year of data)
ITU Internet GGI	78	1.00 (78)	.820 (7)	varies (2012–15)
GSMA Mobile Phone GGI	22	.820 (7)	1.00 (22)	varies (2015)
Internet Penetration	191	.650 (78)	.564 (22)	Annual (2016–17)
GDP per Capita 2016	175	.353 (74)	.551 (22)	Annual (2016)
log(GDP per Capita)	175	.592 (74)	.593 (22)	Annual (2016)
Human Development Index (HDI)	181	.629 (76)	.570 (22)	Annual (2015)
Adult Literacy rate (HDI)	144	.535 (54)	.507 (22)	Annual (2015)
HDI – Education	181	.594 (76)	.595 (22)	Annual (2015)
HDI – Mean Years of Schooling	181	.602 (76)	.528 (22)	Annual (2015)
HDI – Secondary Education Rate	158	.492 (76)	.463 (21)	Annual (2015)
HDI – Unemployment sex ratio	171	−.038 (76)	.066 (22)	Annual (2015)
HDI – Income	181	.581 (76)	.583 (22)	Annual (2015)
OMPI	100	−.765 (22)	−.572 (22)	Biannual (2005–15)
FB penetration	191	.447 (78)	.400 (22)	Real time (2017)
FB/Internet penetration ratio	191	.021 (78)	−.084 (22)	Annual (2016–17)
USP	100	.728 (35)	.672 (21)	Annual (2015–16)
low USP Indicator	100	−.581 (35)	−.767 (21)	Annual (2015–16)
South Asia Indicator	191	−.425 (78)	−.613 (22)	Static (2017)
GGGR Score	139	.297 (75)	.079 (19)	Annual (2015)
GGGR – Economy	139	.380 (75)	.330 (19)	Annual (2015)
GGGR – Labor Force	139	.341 (75)	.238 (19)	Annual (2015)
GGGR – Wages	133	.068 (74)	.124 (19)	Annual (2015)
GGGR – Income	138	.296 (75)	.292 (19)	Annual (2015)
GGGR – Managerial Work	118	.473 (73)	.427 (12)	Annual (2015)
GGGR – Professional work	117	.450 (72)	.699 (12)	Annual (2015)
GGGR – Education	139	.618 (75)	.241 (19)	Annual (2015)
GGGR – Literacy	138	.510 (74)	.348 (19)	Annual (2015)
GGGR – Primary Educ.	126	.042 (67)	.187 (18)	Annual (2015)
GGGR – Secondary Educ.	136	.208 (75)	.160 (19)	Annual (2015)
GGGR – Tertiary Educ.	132	.642 (73)	.141 (19)	Annual (2015)
GGGR – Health	139	.038 (75)	.051 (19)	Annual (2015)
GGGR – Sex Ratio at Birth	139	−.019 (75)	−.024 (19)	Annual (2015)
GGGR – Life Expectancy	139	.055 (75)	.179 (19)	Annual (2015)
GGGR – Politics	139	.086 (75)	−.438 (19)	Annual (2015)
GGGR – Parliament	137	.036 (75)	−.072 (19)	Annual (2015)
GGGR – Ministerial	139	.208 (75)	.103 (19)	Annual (2015)
GGGR – Female Head of State	139	−.044 (75)	−.520 (19)	Annual (2015)
FB GGI age 18+	178	.834 (76)	.650 (22)	Real time (2017)
FB GGI age 15–19	179	.634 (77)	.659 (22)	Real time (2017)
FB GGI age 20–24	179	.757 (77)	.705 (22)	Real time (2017)
FB GGI age 25–29	179	.808 (77)	.739 (22)	Real time (2017)
FB GGI age 30–34	179	.793 (77)	.691 (22)	Real time (2017)
FB GGI age 35–39	179	.777 (77)	.603 (22)	Real time (2017)
FB GGI age 40–44	179	.784 (77)	.566 (22)	Real time (2017)
FB GGI age 45–49	179	.758 (77)	.529 (22)	Real time (2017)
FB GGI age 50–54	179	.742 (77)	.468 (22)	Real time (2017)
FB GGI age 55–59	179	.700 (77)	.480 (22)	Real time (2017)
FB GGI age 60–64	179	.674 (77)	.481 (22)	Real time (2017)
FB GGI age 65+	178	.596 (76)	.378 (22)	Real time (2017)
FB GGI age 18–23	178	.724 (76)	.690 (22)	Real time (2017)
FB GGI age 20+	178	.828 (76)	.654 (22)	Real time (2017)
FB GGI age 20–64	178	.818 (76)	.660 (22)	Real time (2017)
FB GGI age 21+	178	.835 (76)	.663 (22)	Real time (2017)
FB GGI age 25+	178	.819 (76)	.631 (22)	Real time (2017)
FB GGI age 25–49	178	.819 (76)	.654 (22)	Real time (2017)
FB GGI age 25–64	178	.804 (76)	.648 (22)	Real time (2017)
FB GGI age 50+	178	.706 (76)	.472 (22)	Real time (2017)
FB GGI age 60+	178	.645 (76)	.413 (22)	Real time (2017)
FB android device GGI	178	.710 (76)	.616 (22)	Real time (2017)
FB iOS device GGI	178	.815 (76)	.481 (22)	Real time (2017)
FB mobile device GGI	178	.803 (76)	.639 (22)	Real time (2017)
FB feature phone GGI	159	.423 (74)	.444 (22)	Real time (2017)
FB iPhone 7 GGI	173	.675 (76)	.400 (22)	Real time (2017)
FB Smart Phone GGI	178	.792 (76)	.637 (22)	Real time (2017)

Appendix C. Availability of internet and mobile gender gap index by income and development status country classifications

Table 10
Breakdown by income of the number of countries with ground truth Mobile phone and Internet GGI from ITU (ITU, 2016) and GSMA (GSMA Intelligence, 2015a) data as well as number of countries for which a GGI can be predicted with the Online model using just the Facebook GGI. The number in parenthesis indicates the additional number of countries which do not have the ground truth data for which prediction can be made.

	World Bank Income Classification				
Countries with:	Low	Lower-Middle	Upper-Middle	High	NA[a]
in the world	31	53	56	78	–
in dataset	29	50	53	61	–
internet GGI (ITU)	2	11	24	45	2
Online Model prediction	24 (+22)	40 (+30)	40 (+18)	48 (+6)	–
mobile Phone GGI (GSMA)	8	9	5	0	–
Online Model prediction	24 (+17)	40 (+31)	40 (+36)	49 (+49)	–

[a] Palestine and Montserrat did not appear in the World Bank list of countries classified according to Income.

Table 11
Breakdown by development status of the number of countries ground truth Mobile phone and Internet GGI from ITU (ITU, 2016) and GSMA (GSMA Intelligence, 2015a) data as well as number of countries for which a GGI can be predicted with the Online model using just the Facebook GGI. The number in parenthesis indicates the additional number of countries which do not have the ground truth data for which prediction was made.

	UN Development Status Classification			
Countries with:	Least Developed	Developing[a]	Developed	NA[b]
in the world	48	136	64	–
in dataset	46	97	48	2
internet GGI (ITU)	3	41	39	1
Online Model prediction	34 (+31)	74 (+40)	44 (+5)	–
mobile Phone GGI (GSMA)	9	13	0	–
Online Model prediction	34 (+26)	75 (+63)	44 (+44)	–

[a] In the UN Classification list, "Least Developed" countries are a subset of "Developing" countries. To avoid double counting, here the title "Developing" actually refers to countries that are "Developing" AND NOT "Least Developed".
[b] Countries for which a classification was not available. From the countries in the dataset the two countries/territories of Kosovo and Taiwan could not be classified as they did not appear in the UN list of countries by Development status.

Appendix D. Further details about Facebook data

The Facebook data collected for the purposes of this study come from Facebook's Marketing Application Programming Interface (API).[15] This is a tool provided to advertisers to manage functionality related to advertising and managing ads placed through Facebook. The particular method used was the "Ad Account Delivery Estimate" which returns a range of information, including an estimate of "monthly active users" (MAU) on Facebook matching specified targeting specifications (e.g. age, gender, geographical attributes).[16] The MAU counts supplied by the API serve as the user estimates that we use to compute the gender gap indicators in the main text of the paper. The data are publicly-accessible for anyone with a Facebook advertising account.

In order to collect this data we used a Python wrapper library.[17] that provides easy automation for obtaining these estimates. In particular the library was provided with a list of specific targeting criteria for which it then collected the user estimates used in this study. These targeting criteria included the list of countries, gender, various age groups and device types. For each query matching a set of targeting criteria, a number, e.g 40,000, is returned. We use these aggregate numbers in our analysis.

[15] https://developers.facebook.com/docs/marketing-apis.
[16] https://developers.facebook.com/docs/marketing-api/reference/ad-account/delivery_estimate.
[17] https://github.com/maraujo/pysocialwatcher. Details about how to use this library are available on the link provided.

References

Abu-Ghaida, D., & Klasen, S. (2004). The costs of missing the millennium development goal on gender equity. *World Development, 32*(7), 1075–1107.

Abu-Shanab, E., & Al-Jamal, N. (2015). Exploring the gender digital divide in Jordan. *Gender, Technology and Development, 19*(1), 91–113. https://doi.org/10.1177/0971852414563201. Retrieved from https://doi.org/10.1177/0971852414563201.

Alkire, S., & Robles, G. (2017). *Multidimensional poverty index summer 2017: Brief methodological note and results*. University of Oxford. Retrieved 2017-07-26, from http://www.ophi.org.uk/wp-content/uploads/OPHIMethNote_44_Summer_2017.pdf.

Alozie, N. O., & Akpan-Obong, P. (2017). The digital gender divide: Confronting obstacles to women's development in Africa. *Development Policy Review, 35*(2), 137–160. https://doi.org/10.1111/dpr.12204. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/dpr.12204/abstract.

Antonio, A., & Tuffley, D. (2014). The gender digital divide in developing countries. *Future Internet, 6*(4), 673–687. https://doi.org/10.3390/6040673. Retrieved 2017-07-16, from http://www.mdpi.com/1999-5903/6/4/673.

Anzia, S., & Berry, C. (2011). The Jackie (and Jill) Robinson effect: Why do congresswomen outperform congressmen? *American Journal of Political Science, 55*(3), 478–493.

Araújo, M., Mejova, Y., Weber, I., & Benevenuto, F. (2017). Using facebook ads audiences for global lifestyle disease surveillance: Promises and limitations. In *Websci* (pp. 253–257).

Bimber, B. (2000). Measuring the gender gap on the internet. *Social Science Quarterly, 81*(3), 868–876. Retrieved from http://www.jstor.org/stable/42864010.

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science, 350*(6264), 1073–1076.

Boyd, D., & Ellison, N. (2010). Social network sites: Definition, history, and scholarship. *IEEE Engineering Management Review, 3*(38), 16–31.

Brännström, I. (2012). Gender and digital divide 2000–2008 in two low-income economies in sub-saharan africa: Kenya and somalia in official statistics. *Government Information Quarterly, 29*(1), 60–67.

Broadband Commission. (2013). *Doubling Digital Opportunities – Enhancing the Inclusion of Women and Girls in the Information Society* (Tech. Rep.). UNESCO;

ITU. Retrieved 2017-07-11, from http://www.broadbandcommission.org/Documents/working-groups/bb-doubling-digital-2013.pdf.
Chunara, R., Bouton, L., Ayers, J. W., & Brownstein, J. S. (2013). Assessing the online social environment for surveillance of obesity prevalence. *PLoS One, 8*(4), 1–8.
Cooper, J. (2006). The digital divide: The special case of gender. *Journal of Computer Assisted Learning, 22*(5), 320–334. https://doi.org/10.1111/j.1365-2729.2006.00185.x. Retrieved 2017-07-31, from http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2729.2006.00185.x/abstract.
di Bella, E., Leporatti, L., & Maggino, F. (2018). Big data and social indicators: Actual trends and new perspectives. *Social Indicators Research*, 1–10.
DiMaggio, P., Hargittai, E., Celeste, C., & Shafer, S. (2004). *From unequal access to differentiated use: A literature review and agenda for research on digital inequality* (Tech. Rep.). Russell Sage Foundation.
DiMaggio, P., Hargittai, E., Neuman, W. R., & Robinson, J. P. (2001). Social implications of the internet. *Annual Review of Sociology, 27*(1), 307–336. Retrieved 2017-07-31, from https://doi.org/10.1146/annurev.soc.27.1.307.
Dunahee, M., & Lebo, H. (2016). *The world internet project international report, 6th edition* (Tech. Rep.). Retrieved from http://www.digitalcenter.org/wp-content/uploads/2013/06/2015-World-Internet-Report.pdf.
Elvidge, C. D., Sutton, P. C., Ghosh, T., Tuttle, B. T., Baugh, K. E., Bhaduri, B., ... Bright, E. (2009). A global poverty map derived from satellite data. *Computers & Geosciences, 35*(8), 1652–1660.
Evangelos, K., Efthimios, T., & Konstantinos, T. (2013). Understanding the predictive power of social media. *Internet Research, 23*(5), 544–559. https://doi.org/10.1108/IntR-06-2012-0114. Retrieved from http://www.emeraldinsight.com/doi/abs/10.1108/IntR-06-2012-0114.
Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1) New York: Springer Series in Statistics.
Ganguli, I., Hausmann, R., & Viarengo, M. (2014). Closing the gender gap in education: What is the state of gaps in labour force participation for women, wives and mothers? *International Labour Review, 153*(2), 173–207.
Giannone, D., Reichlin, L., & Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics, 55*(4), 665–676.
GSMA Intelligence. (2015a). *Bridging the gender gap: Mobile access and usage in low- and middle-income countries – methodology* (Tech. Rep.). GSMA. Retrieved 2017-07-30, from https://www.gsma.com/mobilefordevelopment/wp-content/uploads/2016/03/GSMA_Bridging-the-gender-gap_Methodology3.2015.pdf.
GSMA Intelligence. (2015b). *The mobile economy – Europe 2015* (Tech. Rep.). GSM Association. Retrieved 2017-07-26, from https://www.gsmaintelligence.com/research/?file=06d1c45d0528233e7a9560843d85c8bd&download.
GSMA Intelligence. (2016a). *The mobile economy – Asia Pacific 2016* (Tech. Rep.). GSM Association. Retrieved 2017-07-26, from https://www.gsmaintelligence.com/research/?file=5369cb14451e0db728bd266c7657a251&download.
GSMA Intelligence. (2016b). *The mobile economy – Middle East and North Africa 2016* (Tech. Rep.). GSM Association. Retrieved 2017-07-26, from https://www.gsmaintelligence.com/research/?file=9246bbe14813f73dd85b97a90738c860&download.
GSMA Intelligence. (2017a). *The mobile economy – Asia Pacific 2017* (Tech. Rep.). GSM Association. Retrieved 2017-07-26, from https://www.gsmaintelligence.com/research/?file=336a9db2ab3ed95bc70e62bf7e867855&download.
GSMA Intelligence. (2017b). *The mobile economy – Sub-Saharan Africa 2017* (Tech. Rep.). GSM Association. Retrieved 2017-07-26, from https://www.gsmaintelligence.com/research/?file=7bf3592e6d750144e58d9dcfac6adfab&download.
Gurumurthy, A., & Chami, N. (2014). *Gender equality in the information society* (Tech. Rep.). IT for Change. Retrieved from http://www.itforchange.net/sites/default/files/Final%20Policy%20Brief%20pdf.
Haferkamp, N., Eimler, S. C., Papadakis, A.-M., & Kruck, J. V. (2012). Men are from mars, women are from venus? Examining gender differences in selfpresentation on social networking sites. *Cyberpsychology, Behavior, and Social Networking, 15*(2), 91–98.
Hafkin, N. J., & Huyer, S. (2006). *Cinderella or cyberella?: Empowering women in the knowledge society*. Kumarian Press, Incorporated.
Hafkin, N. J., & Huyer, S. (2007). Women and Gender in ICT Statistics and Indicators for Development. *Information Technologies & International Development, 4*(2), 25–41. Retrieved 2017-08-11, from http://itidjournal.org/index.php/itid/article/view/254.
Haight, M., Quan-Haase, A., & Corbett, B. A. (2014). Revisiting the digital divide in Canada: The impact of demographic factors on access to the internet, level of online activity, and social networking site usage. *Information, Communication & Society, 17*(4), 503–519. https://doi.org/10.1080/1369118X.2014.891633. Retrieved from https://doi.org/10.1080/1369118X.2014.891633.
Harald, S., Daniel, G.-A., Panagiotis, T. M., Eni, M., Markus, S., & Peter, G. (2013). The power of prediction with social media. *Internet Research, 23*(5), 528–543. https://doi.org/10.1108/IntR-06-2013-0115. Retrieved fromhttp://www.emeraldinsight.com/doi/abs/10.1108/IntR-06-2013-0115.
Hargittai, E., & Shafer, S. (2006). Differences in actual and perceived online skills: The role of gender*. *Social Science Quarterly, 87*(2), 432–448. https://doi.org/10.1111/j.1540-6237.2006.00389.x. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1540-6237.2006.00389.x/abstract.
Hilbert, M. (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), 479–489. https://doi.org/10.1016/j.wsif.2011.07.001. Retrieved from http://www.sciencedirect.com/science/article/pii/S0277539511001099.
Huyer, S., & Carr, M. (2002). Information and communication technologies: A priority for women. *Gender, Technology and Development, 6*(1), 85–100.
IEAG. (2014). *A world that counts-mobilising the data revolution for sustainable development*. Report prepared for the UN Secretary General by the Independent Expert Advisory Group on a Data Revolution for Sustainable Development. Retrieved from http://www.undatarevolution.org/wp-content/uploads/2014/11/A-World-That-Counts.pdf.
Intel. (2012). *Women and the Web* (Tech. Rep.). Retrieved 2017-07-31, from http://www.intel.com/content/www/us/en/technology-in-education/women-in-the-web.html.
ITU. (2015). *ICT Facts and Figures 2015* (Tech. Rep.). ITU. Retrieved 2017-07-13, from http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2015.pdf.
ITU. (2016). Individuals using the Internet (from any location), by gender and urban/rural location (%). Retrieved from http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx.
ITU. (2017). *ICT Facts and Figures 2017* (Tech. Rep.). ITU. Retrieved 2017-08-16TZ, from http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2017.pdf.
Joiner, R., Messer, D., Littleton, K., & Light, P. (1996). Gender, computer experience and computer-based problem solving. *Computers & Education, 26*(1), 179–187. https://doi.org/10.1016/0360-1315(96)00008-5. Retrieved 2017-08-11, from http://www.sciencedirect.com/science/article/pii/0360131596000085.
Joinson, A. N. (2008). Looking at, looking up or keeping up with people?: Motives and use of facebook. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1027–1036). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1357054.1357213, https://doi.org/10.1145/1357054.1357213.
Krasnova, H., Veltri, N. F., Eling, N., & Buxmann, P. (2017). Why men and women continue to use social networking sites: The role of gender differences. *The Journal of Strategic Information Systems*.
Lampos, V., & Cristianini, N. (2012). Nowcasting Events from the Social Web with Statistical Learning. *ACM Trans. Intell. Syst. Technol., 3*(4). https://doi.org/10.1145/2337542.2337557. 72:1–72:22. Retrieved from http://doi.acm.org/10.1145/2337542.2337557.
Letouze, E., & Jutting, J. (2014). *Official statistics, big data and human development: towards a new conceptual and operational approach* (Tech. Rep.). Data Pop Alliance and PARIS21. Retrieved 2017-07-30TZ, from https://www.odi.org/sites/odi.org.uk/files/odi-assets/events-documents/5161.pdf.
Magno, G., & Weber, I. (2014). International gender differences and gaps in online social networks. In *Social informatics (socinfo)* (pp. 121–138).
Mao, H., Shuai, X., Ahn, Y.-Y., & Bollen, J. (2015). Quantifying socioeconomic indicators in developing countries from mobile phone communication data: applications to côte d'ivoire. *EPJ Data Science, 4*(1).
Messias, J., Vikatos, P., & Benevenuto, F. (2017). White, man, and highly followed: Gender and race inequalities in twitter. In *Web intelligence* (p. to appear).
Miniwatts Marketing Group. (2017). *World Internet Users Statistics and 2017 World Population Stats*. Retrieved 2017-06-23, from http://www.internetworldstats.com/stats.htm.
Moolman, J., Primo, N., & Shackleton, S.-J. (2007). Introduction: Taking a byte of technology: women and ICTs. *Agenda: Empowering Women for Gender Equity* (71), 4–14. https://doi.org/10.2307/27739232. Retrieved 2017-08-11, from http://www.jstor.org/stable/27739232.
Norris, P. (2001). *Digital divide: Civic engagement, information poverty, and the internet worldwide*. Cambridge University Press.
Ono, H., & Zavodny, M. (2007). Digital inequality: Five country comparison using microdata. *Social Science Research, 36*(3), 1135–1155. https://doi.org/10.1016/j.ssresearch.2006.09.001. Retrieved 2017-07-31, from http://www.sciencedirect.com/science/article/pii/S0049089X0600072X.
Pew Research Centre. (2015, August). Men catch up with women on overall social media use. Retrieved from http://www.pewresearch.org/fact-tank/2015/08/28/men-catch-up-with-women-on-overall-social-media-use/.
Pew Research Centre. (2016, November). Social media update 2016. Retrieved from http://assets.pewresearch.org/wp-content/uploads/sites/14/2016/11/10132827/PI_2016.11.11_Social-Media-Update_FINAL.pdf.
Qiang, C. Z.-W., Clarke, G. R., & Halewood, N. (2006). *Information and communications for development: Global trends and policies*. Washington, DC: World Bank.
Rice, R. E., & Katz, J. E. (2003). Comparing internet and mobile phone usage: Digital divides of usage, adoption, and dropouts. *Telecommunications Policy, 27*(8), 597–623. https://doi.org/10.1016/S0308-5961(03)00068-5. Retrieved 2017-08-11, from http://www.sciencedirect.com/science/article/pii/S0308596103000685.
Rogers, E. M. (2010). *Diffusion of innovations*. Simon and Schuster.
Saha, K., Weber, I., Birnbaum, L. M., & De Choudhury, M. (2017). Characterizing awareness of schizophrenia among facebook users by leveraging facebook advertisement estimates. *Journal of Medical Internet Research, 19*(5), e156.
Santosham, S., & Lindsey, D. (2015). *Bridging the gender gap: Mobile access and usage in low- and middle-income countries* (Tech. Rep.). GSMA Intelligence. Retrieved 2017-07-27, from https://www.gsma.com/mobilefordevelopment/wp-content/uploads/2016/02/GSM0001_03232015_GSMAReport_NEWGRAYS-Web.pdf.
Sweeny, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10*(05), 557–570.

United Nations Population Division. (2017). *World population prospects: The 2015 revision*. Retrieved 2017-08-19TZ, from https://esa.un.org/unpd/wpp/Download/Standard/Population/.

Unwin, P. T. H. (2009). *ICT4d: Information and communication technology for development*. Cambridge University Press.

Wagner, C., Graells-Garrido, E., Garcia, D., & Menczer, F. (2016). Women through the glass ceiling: gender asymmetries in wikipedia. *EPJ Data Science, 5*(1), 5.

Walsham, G., & Sahay, S. (2006). Research on information systems in developing countries: Current landscape and future prospects. *Information Technology for Development, 12*(1), 7–24.

Wasserman, I. M., & Richmond-Abbott, M. (2005). Gender and the internet: Causes of variation in access, level, and scope of use. *Social Science Quarterly, 86*(1), 252–270. https://doi.org/10.1111/j.0038-4941.2005.00301.x. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.0038-4941.2005.00301.x/abstract.

World Bank. (2017). *GDP per capita, PPP (current international $) | Data*. Retrieved 2017-08-19, fromhttp://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD.

World Economic Forum. (2016). *Global Gender Gap Report 2016* (Tech. Rep.). World Economic Forum. Retrieved 2017-07-17, from http://wef.ch/1yrt8iq.

WWW Foundation. (2015). *Women's Rights Online: Translating Access into Empowerment*. Retrieved 2017-08-16, from https://webfoundation.org/research/womens-rights-online-2015/.

Yazdani, M., & Manovich, L. (2015). Predicting social trends from nonphotographic images on twitter. In *2015 IEEE international conference on big data (big data)* (pp. 1653–1660).

Zagheni, E., & Weber, I. (2012). You are where you e-mail: Using e-mail data to estimate international migration rates. In *Proceedings of the 4th annual ACM web science conference* (pp. 348–351). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2380718.2380764, https://doi.org/10.1145/2380718.2380764.

Zagheni, E., Weber, I., & Gummadi, K. (2017). Leveraging facebook's advertising platform to monitor stocks of migrants. *Population and Development Review, 43*(4), 721–734. https://doi.org/10.1111/padr.12102.