Analyzing gender inequality through large-scale Facebook advertising data

David Garcia[a,b], **Yonas Mitike Kassa**[c,d], **Angel Cuevas**[d], **Manuel Cebrian**[e,f], **Esteban Moro**[f,g], **Iyad Rahwan**[f,h,1], **and Ruben Cuevas**[d,1]

[a]Complexity Science Hub Vienna, 1080 Vienna, Austria; [b]Section for Science of Complex Systems, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, 1090 Vienna, Austria; [c]IMDEA Networks Institute, 28918 Leganés, Spain; [d]Department of Telematic Engineering, Universidad Carlos III de Madrid, 28911 Leganés, Spain; [e]Data61, Commonwealth Scientific and Industrial Research Organisation, 3008 Melbourne, Australia; [f]The Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139; [g]Grupo Interdisciplinar de Sistemas Complejos, Department of Mathematics, Universidad Carlos III de Madrid, 28911 Leganes, Spain; and [h]Institute for Data, Systems & Society, Massachusetts Institute of Technology, Cambridge, MA 02139

Edited by Michael Macy, Cornell University, Ithaca, NY, and accepted by Editorial Board Member Mary C. Waters May 5, 2018 (received for review October 14, 2017)

Online social media are information resources that can have a transformative power in society. While the Web was envisioned as an equalizing force that allows everyone to access information, the digital divide prevents large amounts of people from being present online. Online social media, in particular, are prone to gender inequality, an important issue given the link between social media use and employment. Understanding gender inequality in social media is a challenging task due to the necessity of data sources that can provide large-scale measurements across multiple countries. Here, we show how the Facebook Gender Divide (FGD), a metric based on aggregated statistics of more than 1.4 billion users in 217 countries, explains various aspects of worldwide gender inequality. Our analysis shows that the FGD encodes gender equality indices in education, health, and economic opportunity. We find gender differences in network externalities that suggest that using social media has an added value for women. Furthermore, we find that low values of the FGD are associated with increases in economic gender equality. Our results suggest that online social networks, while suffering evident gender imbalance, may lower the barriers that women have to access to informational resources and help to narrow the economic gender gap.

gender divide | Facebook | social media | inequality | development

The Web was designed to be universally accessible and open, carrying the promise of equal opportunity in the access to online information and services (1) as the great potential equalizer (2). However, despite the widespread adoption of the Web and other information communication technologies (ICTs), online access is heterogeneously distributed across demographic factors, such as income and gender—a phenomenon called the digital divide (3–5).

Governments and global organizations express their concern about the digital divide, aiming to connect the 4 billion people that remain offline (6, 7). However, the effects of increasing Internet penetration in development are rarely backed up against empirical data (8), and the latest report by the World Bank suggests that unequally distributed growth in Internet penetration might exacerbate socioeconomic inequalities (7). Beyond the divide in the access to the Internet, there are further challenges with respect to digital inequality: the heterogeneity of online activity and engagement across demographic groups (2).

Among online resources, social media play a key role in economic development (for example, by providing information that facilitates finding employment) (9, 10). An important open question is whether equality in the access to social media can work as a digital provide (11), bringing equality in other social, political, and economic aspects of society. The World Wide Web Foundation reports that one of the key elements in the digital divide is gender inequality (12). Social media data show the traces of gender inequalities from content biases and activity on Wikipedia (13, 14) to visibility and interaction disparities on Twitter (15–17) and professional gender gaps in LinkedIn (18). Empirical analyses of digital traces have the potential to track more general demographic patterns (19, 20), such as fertility rates (21).

To properly understand the digital divide, a pervasive problem in cross-country comparisons is the limited size of country samples and the challenges to generate unbiased survey data (7). To overcome this issue, we deployed a system to collect large-scale data from the Facebook online social network through its marketing Application Programming Interface (API), as explained more in detail in *Materials and Methods*. The Facebook marketing API has been useful in previous research to estimate the value of user data (22), to approximate the size and integration of migrant populations (23–25), and to generate estimations of Internet and mobile phone gender gaps that explain 69% of the variance of International Telecommunications Union measurements (20). For our study of the relationship between social media gender divides and other economic, education, heath, and political gender inequalities, we generated an anonymous dataset with statistics about the total number of registered users and daily active users (DAUs) of each gender in each country. While

Significance

We present the Facebook Gender Divide, an inexpensive, real-time instrument for measuring gender differences in Facebook access and activity in 217 countries. The Facebook Gender Divide captures standard indicators of Internet penetration and gender equality indices in education, health, and economic opportunity. We find that the tendency of countries to approach economic gender equality is negatively associated with a high Facebook Gender Divide. Our results suggest that online social networks, while suffering gender imbalance, may lower information access barriers for women and narrow the economic gender gap.

Author contributions: D.G., A.C., E.M., and R.C. designed research; D.G. and Y.M.K. performed research; D.G., Y.M.K., A.C., and R.C. analyzed data; Y.M.K., A.C., M.C., E.M., I.R., and R.C. contributed to writing; and D.G., A.C., M.C., I.R., and R.C. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. M.M. is a guest editor invited by the Editorial Board.

Data deposition: The data reported in this paper are available at https://github.com/dgarcia-eu/FacebookGenderDivide.

[1] To whom correspondence may be addressed. Email: irahwan@mit.edu or rcuevas@it.uc3m.es.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1717781115/-/DCSupplemental.

our dataset does not contain personal information on any individual user, our study covers a total of 217 countries and more than 1.4 billion users.

Our dataset allows for the quantification of Facebook activity ratios of each gender in each country. From them, we calculate the Facebook Gender Divide (FGD) as the logarithm of the ratio between the activity ratios for men and for women (more details are in *Materials and Methods*). The FGD has a value below zero when women tend to be more active on Facebook that men, a value close to zero for equal activity tendencies, and a positive value when men are more active on Facebook than women in a country. Our computation of the FGD is consistent with similar measurements constructed from limited survey samples from the Pew Research Center and the Global Web Index (GWI) as we comment in *Materials and Methods* and show in *SI Appendix*, section 1.

Furthermore, the Facebook marketing API allows us to make precise estimates of the Facebook penetration in a country calculated as the total number of user accounts (independent of gender and activity) over the total population of the country. We combine these measurements with standard socioeconomic indices, including gross domestic product, Internet penetration, and economic inequality, as well as indices from the World Economic Forum Gender Gap Report that measure gender equality in terms of education, health, political participation, and economic opportunities (26).

Results

Fig. 1 shows a world map with countries colored according to their FGD, revealing that many countries are very close to gender equality in Facebook (blue color in Fig. 1). The red scale in Fig. 1 shows countries with positive FGD—that is, a higher proportion of males on Facebook. The range of values toward FGD below zero (more tendency for women to be on Facebook) is much narrower than above zero as can be seen in the scatterplot with the activity ratios of each gender (Fig. 1, *Left Inset*) and in the skewness of the distribution of FGD across countries (Fig. 1, *Right Inset*).

Countries with high FGD are located around Africa and southwest Asia, as shown in Fig. 1. This suggests that variations in socioeconomic factors of gender inequality across regions could be explanatory of the FGD. We test this observation using a linear regression model of the FGD as a function of the four indices of gender equality measured by the World Economic Forum (economic opportunity, education, health, and political participation) plus five nongender-based controls of Internet penetration, population size, economic inequality, Facebook penetration, and mean Facebook active user age (see *Materials and Methods*). Fig. 2*A* shows the quality of the model fit, comparing empirical values of FGD rank vs. model predictions. Remarkably, the model can explain well the ranking of FGD ($R^2 = 0.74$), with very few points far from the diagonal. While this result might be partially explained by Facebook using vital statistics in their calculations, it is nevertheless consistent with replications of the model using limited survey samples from the Pew Research Center and the GWI (*SI Appendix*, section 2). This indicates that the performance of the model is not an artifact of the Facebook marketing API.

Fig. 2*B* shows the estimate of the coefficients of our model of FGD. The strongest coefficient is that of education gender equality, which can also be observed from the colors in Fig. 2*A*. Specifically, countries with high rank in this index have, on average, lower FGD. Health and economic gender equality also have significant negative coefficient estimates, showing that the FGD captures more than one type of inequality. Note that the index for political gender equality does not have a significant relationship with FGD when the other indices are considered in the model.

Among gender-independent controls, only Internet penetration is negatively associated with FGD. Nevertheless, the FGD is also correlated with gross domestic product per capita (Spearman correlation -0.57, $p < 10^{-6}$). For that reason, we repeated the model using gross domestic product as a control variable, finding similar results. These results evidence that the relationship between gender equality indices and FGD is observable when development metrics are considered. We present these additional controls, regression diagnostics, and robustness tests in *SI Appendix*, section 2, concluding that the negative relationships between FGD and gender equality indices are robust.

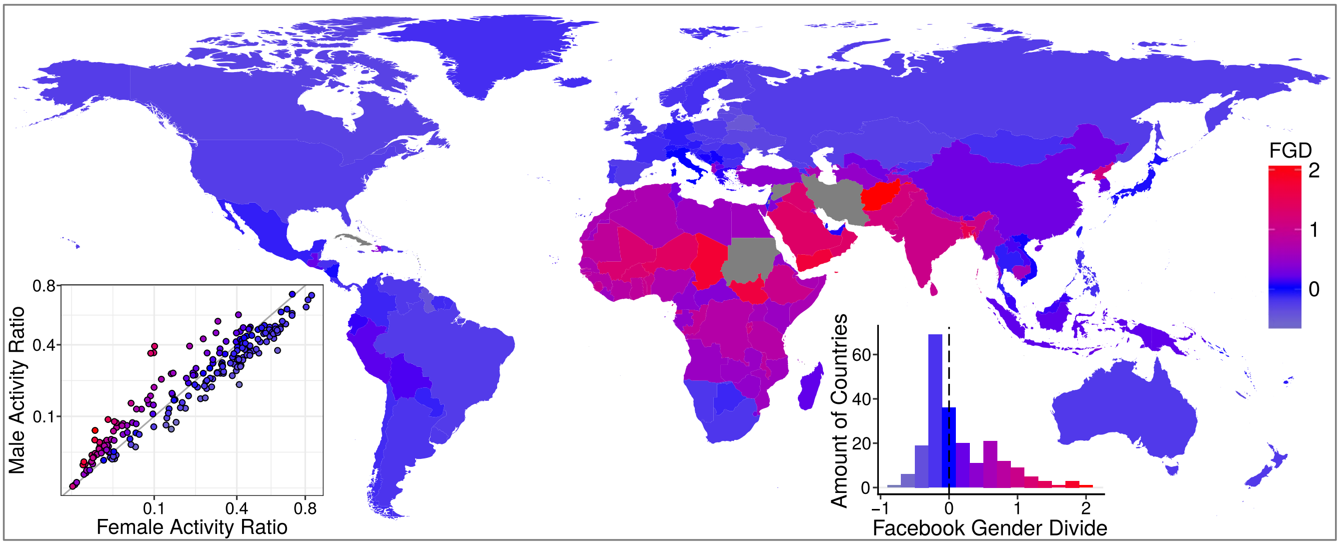

Fig. 1. The FGD across 217 countries. Countries are colored according their FGD from highly skewed toward males (red) and balanced (blue) to highly skewed toward females (green; not visible). *Left Inset* shows the scatterplot of male and female activity ratios across all countries, revealing a spread along the diagonal. *Right Inset* shows the histogram of FGD values in bins of width 0.2. While the mode of countries is slightly below zero, there is significant skewness toward high FGD values. An online interactive version of this figure can be found at https://dgarcia-eu.github.io/FacebookGenderDivide/Visualization.html.

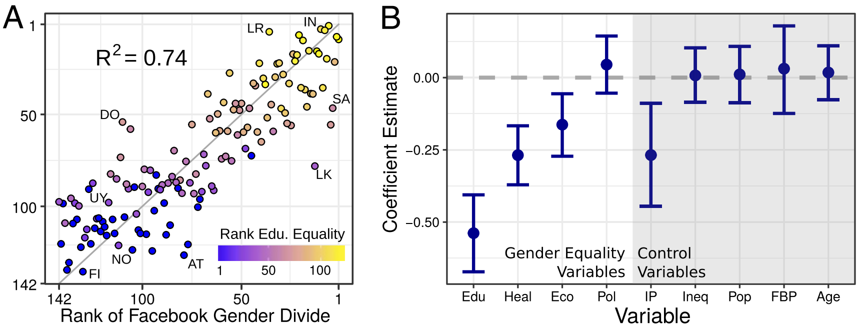

Fig. 2. Regression results of FGD as a function of gender equality. (*A*) Model predictions vs. rank of FGD, where rank 1 is the country with the highest FGD. The model achieves a high R^2 above 0.74, explaining the majority of the variance of the FGD ranking. Some countries are labeled, from high FGD [Liberia (LR), India (IN), and Saudi Arabia (SA)] to low FGD [Finland (FI), Norway (NO), and Uruguay (UY)], as well as some outliers [Dominican Republic (DO), Austria (AT), and Sri Lanka (LK)]. (*B*) Coefficient estimates and 95% CIs of the terms of the regression fit (excluding intercept). Education (Edu), health (Heal), and economic gender equality (Eco) are significantly and negatively associated with the FGD, but political gender equality (Pol) is not. From the control variables, Internet penetration (IP) is negatively associated with FGD, but the rest are not. The main role of education equality in FGD can be observed in *A*, where dots are colored according to the rank of education gender equality, showing that countries with low FGD are ranked high on education gender equality. An online interactive version of this figure can be found in https://dgarcia-eu.github.io/FacebookGenderDivide/Visualization.html. FBP, Facebook penetration; Ineq, income inequality; Pop, total population.

The value of being active on social media might vary across genders, which we address in a wide country comparison. The general penetration of a communication channel can increase the value that individuals get for using it, which is an example of a feedback mechanism driven by (positive) network externalities (27), also known as Metcalfe's law (28). If there are network externalities on Facebook, the activity ratio of countries should scale superlinearly with the Facebook penetration in each country. This scaling relationship with Facebook penetration might vary for the activity ratios of different genders, which would signal an additional marginal benefit of using Facebook for one gender.

Fig. 3 shows the scaling relationship per gender between the activity ratio and the total Facebook penetration in each country. Lines show the result of a power law fit between both variables with an intercept and an interaction term for gender. The estimate of the scaling exponent for each gender is clearly above one for both genders, revealing a superlinear trend consistent with network externalities in Facebook. This exponent is significantly stronger for female users ($\alpha_F = 1.45$, CI = $[1.41, 1.49]$) than for male users ($\alpha_M = 1.20$, CI = $[1.16, 1.24]$) (details are in *SI Appendix*, section 3), suggesting that the network externalities in Facebook are stronger for women than for men.

Given the network externalities shown above, could the FGD be related to changes in economic gender inequality? We test this possibility by analyzing the change in FGD and economic gender equality between 2015 and 2016. We fitted two regression models, one of changes of economic gender equality as a function of FGD ($FGD_{2015} \rightarrow \Delta Eco_{2016}$) and the converse one ($Eco_{2015} \rightarrow \Delta FGD_{2016}$), including controls for autocorrelation and gross domestic product as explained in *Materials and Methods*. The coefficient estimates, shown in Fig. 4, reveal a significant positive relationship between the FGD rank and changes in economic gender inequality but not vice versa: there is no significant relationship between economic gender equality and the changes in FGD.

The partial R^2 value of FGD_{2015} in the first model is much higher than the equivalent of Eco_{2015} in the second model (median bootstrap values of 0.027 and 0.002, respectively), as shown in Fig. 4, *Right*. This suggests the existence of an association between FGD and changes in economic gender equality, such that countries with a low value of FGD (i.e., high rank number) tend more, on average, to approaching economic gender equality. This observation is consistent across age groups and is robust to the inclusion of further control variables, including socioeconomic indicators, other gender equality metrics, and Hofstede's culture values (29) (more details are in *SI Appendix*, section 4). On the contrary, this association is not observable for education gender inequality, as a model of ΔEdu_{2016} shows no significant coefficient for FGD_{2015}.

Discussion

By quantifying the FGD among 1.4 billion Facebook users, we show a number of phenomena that deserve further investigation. The FGD is associated with other types of gender inequality, including economic, health, and education inequality. While the mechanisms behind this connection and its generalizability to other social media remain open questions, this work is an example of how publicly accessible social media data can be used to understand an important social phenomenon.

Recent reports warn about the possibility that individual Facebook user data were misused by Cambridge Analytica (30), pointing to general concerns about privacy in social media. We share those concerns, in particular with respect to the use of sensitive data in potential conflict with the European Union General

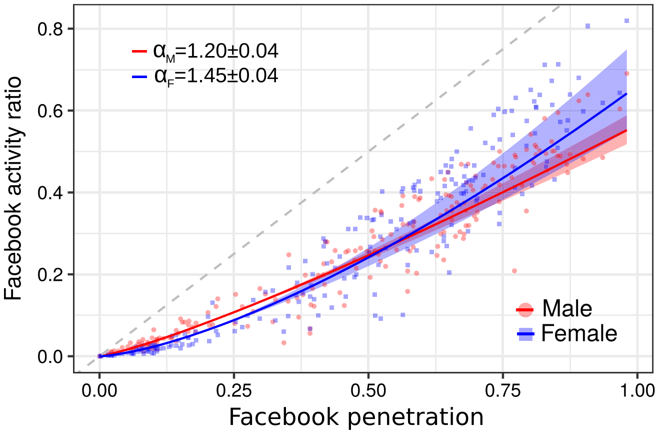

Fig. 3. Gender differences in network externalities on Facebook. Scaling of the Facebook activity ratio per gender vs. total Facebook penetration. Solid lines show fit results, and shaded areas show their 95% CIs. Both male and female activity ratios grow superlinearly with Facebook penetration ($\alpha > 1$), indicating positive network externalities. These network externalities are stronger for female than for male users ($\alpha_F > \alpha_M$).

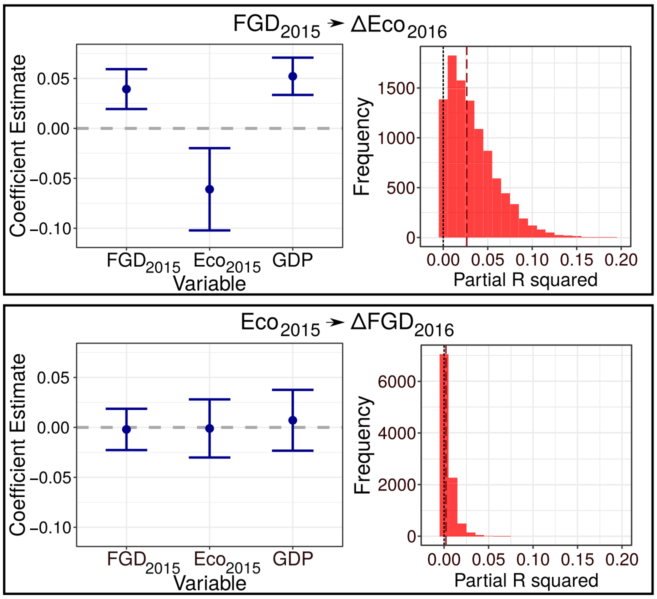

Fig. 4. Analysis of changes in economic gender equality and FGD. Coefficient estimates of the regression model of changes in economic gender equality as a function of FGD and control terms (excluding intercept; *Upper Left*) and of the model of changes in FGD as a function of economic gender equality and control terms (excluding intercept; *Lower Left*). *Right* shows the bootstrap distributions of partial R^2 of FGD_{2015} in the first model and of Eco_{2015} in the second one, with dashed vertical lines showing the median R^2 values: 0.027 (*Upper Right*) in the first model and 0.002 (*Lower Right*) in the second one. The FGD explains changes in economic gender equality much better than economic gender equality explains changes in the FGD. GDP, gross domestic product.

Data Protection Regulation (31) and regarding the possible construction of shadow profiles of nonusers (32, 33). Nevertheless, our results show that nonpersonal data (e.g., anonymous and aggregated data produced by billions of Facebook users) can be used for social good, in particular to understand the issue of gender inequalities in society at large.

We found evidence of gender-dependent network externalities—that is, women might receive higher marginal benefit than men from the general adoption of Facebook in a country. While we only observe traces of this phenomenon at an aggregate level, in the differences between activity rates across countries, these results point toward a unique research direction: using observational data to understand the value of social networking sites across demographic attributes.

The FGD provides an inexpensive and accessible way to compute gender divides in social media that can be tracked over time and across the vast majority of countries. This allowed us to identify a relationship between the FGD in 2015 and changes in economic gender inequality in 2016. This relationship could be produced by three mechanisms: (*i*) a causation path between the FGD and changes in economic gender equality, (*ii*) a more complex causation from economic gender equality on changes in FGD, or (*iii*) the prevalence of a third factor of cultural gender norms that drive both the FGD and economic gender equality. While we find evidence for the first explanation, we must note that the real interplay between the FGD and economic gender equality is probably a combination of all three mechanisms, and only future research with more detailed data can answer how.

Our results show trends across a wide range of countries, but caution should be taken when extrapolating to the future or when predicting about individual countries. Before doing so, we need longitudinal models of changes in development factors in a wide range of countries to find the role of the FGD in broader development sequences that include economic development, health, education, and inequality (34) before formulating policy suggestions. Nevertheless, our results allow us to speculate that social media can be an equalizing force that counteracts other barriers [e.g., those that limit women's mobility (35)] by providing access to greater economic opportunities and social capital. In a similar way as mobile phones increased the life quality of fishermen in India (11), social media might work as a digital provide that helps disfavored groups, despite the still generalized inequalities in access to ICTs and in adoption of social media technologies.

Materials and Methods

The Facebook Global Dataset. We collected the number of Facebook users by age and gender in each country using the Facebook marketing API (36). Among other services, this API delivers data for its commercial customers to provide targeted advertising. When supplied with a specific target population, the API returns the total audience size and the price to reach that target audience through Facebook. We iterated each combination of age and gender values, retrieving the total number of users and the number of DAUs for each segment in each country. Our dataset contains the number of male and female registered users and DAUs for all available countries (the API does not deliver data for certain countries; e.g., Syria, Iran, and Cuba). After removing entries of small countries with missing values or low resolution, our dataset contains the total number of users and DAUs segmented by age and gender for 217 countries. Age data in the API start at 13 y old, increasing by 1 y up to a last bin that contains all users age 65 y old or older. We distribute a dataset to allow for the replication and extension of our results through a Github repository (https://github.com/dgarcia-eu/FacebookGenderDivide).

Ethical Considerations. Our analysis of data from the Facebook marketing API only includes aggregated public information. Although the sample includes data from underage Facebook users, we had no access to any personal identifiable information of any user, and we did not interact or manipulate any research subject. The data retrieval was performed as part of the TYPES (towards transparency and privacy in the online advertising business) Project funded by the European Comission (GA-653449) and was approved by the Committee of Ethics in Research of the Carlos III University of Madrid (Ethics Report CEI-2015-001). Our analysis of the data, in line with the growing consensus in ethics (37), is exempted from ethics review as agreed on by the board of IMDEA Networks and by the executive office of the Complexity Science Hub Vienna. Nevertheless, following the guidelines of the Association of Internet Researchers (38), we consider the possible downstream consequences of our large-scale research. The resolution of the Facebook marketing API prevents the singling out of individual users, which makes all our codes useless for identifying individuals of any minority or threatened group. In addition, there is no way to identify the accounts of users and use our analysis for any kind of personalization or individual manipulation. From the onset, our project had the potential to reveal important relationships between social media use and gender inequalities online and offline. These benefits greatly outweigh the minimum risks of analyzing this kind of aggregated data that are accessible to anyone with an Internet connection.

Validating the FGD. Facebook provided the raw data for our study as aggregated values, but as with any research method, we should not take it at face value without comparing it with more established methods. This is of special importance given the challenges previously found with health-related data from this API (39).

To validate our measurements, we use three reference survey datasets: the Global and Internet & Technology Surveys of the Pew Research Center (www.pewglobal.org/dataset/spring-2016-survey-data and www.pewinternet.org/dataset/march-2016-libraries) and the survey of the GWI (https://www.globalwebindex.com). These datasets allow us to compute reference measurements of Facebook penetration and FGD for small samples of countries to be compared with our calculation of the FGD through the Facebook marketing API.

The results of this validation exercise are reported in detail in *SI Appendix*, section 1. We find high correlation coefficients between our measurement of penetration and the equivalents in the GWI and the Pew Global Survey and for the case of FGD as well. These correlations are as good as the correlations between survey datasets, showing that the Facebook API data have comparable quality but a much higher coverage in terms of countries and better temporal resolution. We find low and nonsignificant correlations between the absolute difference between our measurement of FGD and the one from surveys, but nevertheless, we add a control for Facebook penetration in our models to make sure that our results are not an artifact of correlated errors in the quantification of FGD.

We further compare Facebook penetration across age groups in the United States through the Pew Internet & Technology Survey and the GWI. We find very high correlations between age-dependent measurements. In addition, we explore how representative the FGD is for gender divides in other social media as captured by the GWI survey. We found moderate yet significant correlations with other media, such as WhatsApp, Twitter, and YouTube. This shows that, while we should not take Facebook as representative for all social media, there is certain similarity in gender differences that can motivate future research.

Finally, we test for intraday oscillations of the measurement of FGD and Facebook penetration and found extremely consistent values. For the case of the FGD and the network externalities model, we also repeat our analysis on monthly snapshots of Facebook data for a period of 12 mo between 2015 and 2016, calculating median DAU values each month. This way, we can confirm the robustness of our analysis to possible temporal changes in the way that Facebook reports data through their API.

Gender Equality and Development Datasets. To normalize the number of active users over the total population of each country, we use the data collected by the US Census Bureau International Data Base (https://www.census.gov/programs-surveys/international-programs/about/idb.html). This dataset contains estimates of the resident population by age and gender for more than 226 countries. We combine these data with gender equality indices measured by the World Economic Forum Gender Gap reports of 2015 and 2016 (26). This dataset quantifies the magnitude of gender equality in 145 countries, measuring it with respect to four key areas: health, education, economic opportunity, and politics. This report updates the values for education, economic, and political gender equality on a yearly basis, allowing us to measure changes between 2015 and 2016. To account for additional economic and development indicators, we include data from the World Bank and the Human Development Index (40), measuring control variables of gross domestic product at purchasing power parity per capita in 2012, economic inequality as the quintile ratio, and Internet penetration.

Computing the FGD. We quantify the FGD as a comparison of the rates of activity between genders. The DAU measures how many users have logged into Facebook on a given day, which could be either through a web browser or a mobile application. We use the segmented data from 13 to 65 y olds to normalize the DAU over the total population of a country in those ages, truncating all data that are not included in that age range. This way, we avoid introducing a bias with life expectancy and average age. We calculate a stable estimation of the DAU as the median daily value over the month of July 2015, replicating over other months afterward. This way, for each country c and gender $g \in [Female, Male]$ (for simplicity, we take gender as birth sex; i.e., male or female), we have a measurement of the number of active users $A_{g,c}$ between 13 and 65 y old. Additionally, this allows us to calculate the mean user age for a country to include it in our models.

Using the US Census Bureau data, we calculate the total population of each gender between the ages of 13 and 65 y old in each country, which we denote as $P_{g,c}$. This way, we can normalize the total activity in Facebook over the population in the same age ranges, calculating the activity ratios $R_{g,c} = A_{g,c}/P_{g,c}$. We define the FGD in country c as

$$FGD_c = \log\left(\frac{R_{Male,c}}{R_{Female,c}}\right),$$

which compares male and female Facebook activity rates over the population of country c. A country with positive FGD will have a tendency for men to be more present on Facebook, while a country with negative FGD will show the opposite tendency. A country with $FGD = 0$ will have complete equality in the activity tendencies of both genders.

We further compute the Facebook penetration as the ratio between user accounts between 13 and 65 y olds reported by the API (regardless of activity and gender) and the total population of the country between those ages.

Regression Models. We model dependencies between gender equality indicators and the FGD as linear models after applying a rank transformation to all variables, such that rank 1 is the highest possible value of the variable. This way, we explore monotonic dependencies that do not need to be linear. We define this FGD model as

$$FGD = a_f \cdot Q + b_f \cdot C + c_f + \epsilon,$$

where Q is a matrix with the ranks of economic, health, education, and political gender equality in each country and C contains control variables, such as Internet penetration, income inequality, total population, Facebook penetration, and mean user age (Age). c_f is the intercept, and ϵ denotes the residuals as the normally distributed, uncorrelated error of the model.

We analyze the relationship between changes and levels in economic gender equality and of the FGD with two models. First is an equality changes model:

$$\Delta Eco_{2016} = a_o \cdot Eco_{2015} + b_o \cdot FGD_{2015} + c_o \cdot O + d_o + \psi_o.$$

Second is an FGD changes model:

$$\Delta FGD_{2016} = a_q \cdot FGD_{2015} + b_q \cdot Eco_{2015} + c_q \cdot O + d_q + \psi_q,$$

where ΔEco_{2016} and ΔFGD_{2016} are the changes in economic gender inequality and FGD between 2015 and 2016, respectively. Both models include a control for autocorrelation as a term with the unranked value of the variable in 2015 and a main term of the rescaled ranked value of the other variable. Following previous economics research on Facebook data (10), we include various ranked controls in the matrix O, first with a simple correction for gross domestic product and then, with extensions with other controls as for the FGD model.

We report the coefficient estimates of robust regressors for both models. To compare the effects of one variable with the changes in the other, we first residualize the changes by fitting against all controls. Then, we compute the partial R^2 value of the conditioning variable when fitting the residualized values. To understand the uncertainty of this analysis, we bootstrap over 10.000 samples and report the distribution of R^2 values.

We model network externalities as a power law relationship between the activity ratio of a gender ($R_{g,c}$) and the total Facebook penetration for both genders together (P_c) in a joint model that includes an intercept for gender and interaction with gender. We define in this way the network externalities model as

$$\log(R_{g,c}) = \alpha \cdot \log(P_c) + \beta + \delta_{g,Female}(\alpha_F \cdot \log(P_c) + \beta_F) + \phi,$$

where α measures the scaling relationship between the Facebook presence ratio and the activity ratio of male users, α_F is the difference in that relationship for female users, and ϕ is the residuals. The Kronecker delta function $\delta_{g,Female}$ takes a value of one when $g = Female$ and zero otherwise.

All of the above models do not show relevant multicollinearity when measuring variance inflation factors (41).

We report the fit of the FGD model and the network externalities model with Markov Chain Monte Carlo sampling in JAGS (42). We also fit all models with robust regression (43), reporting the results of the changes models in the text and the rest in *SI Appendix*.

To test the validity of the assumptions of our models after fitting, we verify the normality of residuals through Shapiro–Wilk tests (44) and check that residuals are uncorrelated with fitted values and independent variables. For the case of the network externalities model, we additionally analyze multiplicative residuals to test for the possible role of outliers, as shown in more detail in *SI Appendix*.

ACKNOWLEDGMENTS. We thank the Pew Research Center and the GWI for the access to their data to validate the FGD. D.G. acknowledges funding from the Vienna Science and Technology Fund through Vienna Research Group Grant “Emotional Well-Being in the Digital Society” VRG16-005. Y.M.K. acknowledges funding from H2020 (Horizon 2020) EU Project TYPES Grant 653449. A.C. acknowledges funding from H2020 EU Project SMOOTH (GDPR Compliance Cloud Platform for Micro Enterprises) Grant 768741 and Ramón y Cajal Grant RyC-2015-17732. E.M. acknowledges funding from Ministerio de Economía y Competividad (Spain) through Projects FIS2013-47532-C3-3-P and FIS2016-78904-C3-3-P. R.C. acknowledges funding from H2020 EU project ReCRED (from real-world identities to privacy-preserving and attribute-based credentials for device-centric access control) Grant 653417.

1. Berners-Lee T (2010) Long live the web. *Sci Am* 303:80–85.
2. Hargittai E, Hsieh YP (2013) *Digital Inequality*, ed Dutton W (Oxford Univ Press, Oxford), pp 129–150.
3. Brown R, Barram D, Irving L (1995) *Falling Through the Net: A Survey of the "Have Nots" in Rural and Urban America* (National Telecommunications and Information Administration, Washington, DC).
4. Compaine BM (2001) *The Digital Divide: Facing a Crisis or Creating a Myth?* (MIT Press, Cambridge, MA).
5. Norris P (2001) *Digital Divide: Civic Engagement, Information Poverty, and the Internet Worldwide* (Cambridge Univ Press, Cambridge, UK).
6. United Nations General Assembly (2006) *60 Resolution 252* (United Nations, Geneva).
7. World Bank Group (2016) *World Development Report 2016: Digital Dividends* (World Bank, Washington, DC).
8. Friederici N, Ojanperä S, Graham M (2017) The impact of connectivity in Africa: Grand visions and the mirage of inclusive digital development. *Electron J Inf Syst Dev Countries* 79:1–20.
9. Burke M, Kraut R (2013) Using Facebook after losing a job: Differential benefits of strong and weak ties. *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (ACM, New York), pp 1419–1430.
10. Gee LK, Jones JJ, Fariss CJ, Burke M, Fowler JH (2017) The paradox of weak ties in 55 countries. *J Econ Behav Organ* 133:362–372.
11. Jensen R (2007) The digital provide: Information (technology), market performance, and welfare in the south Indian fisheries sector. *Q J Econ* 122:879–924.
12. World Wide Web Foundation (2015) The web and rising global inequality. Available at http://thewebindex.org/report/. Accessed May 27, 2018.
13. Wagner C, Graells-Garrido E, Garcia D, Menczer F (2016) Women through the glass ceiling: Gender asymmetries in wikipedia. *EPJ Data Sci* 5:5.
14. Rizoiu MA, Xie L, Caetano T, Cebrian M (2016) Evolution of privacy loss in Wikipedia. *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining* (ACM, New York), pp 215–224.
15. Garcia D, Weber I, Garimella VRK (2014) Gender asymmetries in reality and fiction: The Bechdel test of social media. *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media* (AAAI Press, Palo Alto, CA), pp 131–140.
16. Nilizadeh S, et al. (2016) Twitter's glass ceiling: The effect of perceived gender on online visibility. *Proceedings of the Tenth International AAAI Conference on Web and Social Media* (AAAI Press, Palo Alto, CA), pp 289–298.
17. Magno G, Weber I (2014) International gender differences and gaps in online social networks. *Proceedings of the International Conference on Social Informatics* (Springer, Cham, Switzerland), pp 121–138.
18. Haranko K, Zagheni E, Garimella K, Weber I (2018) Professional gender gaps across us cities. arXiv1801.09429.
19. Billari FC, Zagheni E (2017) Big data and population processes: A revolution? *Statistics and Data Science: New Challenges, New Generations*. Proceedings of the Conference of the Italian Statistical Society (SIS) (Firenze Univ Press, Firenze, Italy), pp 167–178.
20. Fatehkia M, Kashyap R, Weber I (2018) Using Facebook ad data to track the global digital gender gap. *World Dev* 107:189–209.
21. Ojala J, Zagheni E, Billari FC, Weber I (2017) Fertility and its meaning: Evidence from search behavior. *Proceedings of the Eleventh International AAAI Conference on Web and Social Media* (AAAI Press, Palo Alto, CA), pp 640–643.
22. González Cabañas J, Cuevas Á, Cuevas R (2017) FDVT: Data valuation tool for Facebook users. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (ACM, New York), pp 3799–3809.
23. Pötzschke S, Braun M (2017) Migrant sampling using facebook advertisements: A case study of polish migrants in four european countries. *Soc Sci Comput Rev* 35:633–653.
24. Zagheni E, Weber I, Gummadi K (2017) Leveraging Facebook's advertising platform to monitor stocks of migrants. *Popul Dev Rev* 43:721–734.
25. Dubois A, Zagheni E, Garimella K, Weber I (2018) Studying migrant assimilation through facebook interests. arXiv 1801.09430.
26. World Economic Forum (2016) World economic forum gender gap report. Available at reports.weforum.org/global-gender-gap-report-2016. Accessed May 27, 2018.
27. Katz ML, Shapiro C (1985) Network externalities, competition, and compatibility. *Am Econ Rev* 75:424–440.
28. Hendler J, Golbeck J (2008) Metcalfe's law, web 2.0, and the semantic web. *Web Semant Sci Serv Agents World Wide Web* 6:14–20.
29. Hofstede G (2003) *Cultures and Organizations: Comparing Values, Behaviors, Institutions, and Organizations Across Nations* (Sage Publications, Thousand Oaks, CA).
30. Ingram D (2018) Factbox: Who is Cambridge Analytica and what did it do? Available at https://reut.rs/2GGNb8F. Acccessed April 23, 2018.
31. Cabañs JG, Cuevas A, Cuevas R (2018) Facebook use of sensitive data for advertising in europe. arXiv 1802.05030.
32. Garcia D (2017) Leaking privacy and shadow profiles in online social networks. *Sci Adv* 3:e1701172.
33. Garcia D, Goel M, Agrawal AK, Kumaraguru P (2018) Collective aspects of privacy in the twitter social network. *EPJ Data Sci* 7:3.
34. Spaiser V, Ranganathan S, Mann RP, Sumpter DJ (2014) The dynamics of democracy, development and cultural values. *PloS One* 9:e97856.
35. Uteng T (2012) Gender and mobility in the developing world. *World Development Report* (World Bank, Washington, DC).
36. Facebook (2018) Facebook ads API. Available at https://developers.facebook.com/docs/graph-api. Accessed May 27, 2018.
37. Metcalf J, Crawford K (2016) Where are human subjects in big data research? the emerging ethics divide. *Big Data Soc* 3:2053951716650211.
38. Ess C, Jones S (2004) Ethical decision-making and Internet research: Recommendations from the aoir ethics working committee. *Readings in Virtual Research Ethics: Issues and Controversies* (IGI Global, Hershey, PA), pp 27–44.
39. Araújo M, Mejova Y, Weber I, Benevenuto F (2017) Using Facebook ads audiences for global lifestyle disease surveillance: Promises and limitations. *Proceedings of the 2017 ACM on Web Science Conference* (ACM, New York), pp 253–257.
40. World Bank (2016) World Bank human developent index. Available at hdr.undp.org/en/content/human-development-index-hdi. Accessed October 10, 2017.
41. Chatterjee S, Hadi AS (2015) *Regression Analysis by Example* (John Wiley & Sons, Hoboken, NJ).
42. Plummer M (2003) JAGS: A program for analysis of Bayesian graphical models using gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, eds Hornik K, Leisch F, Zeileis A. Available at https://www.r-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf. Accessed June 1, 2018.
43. Koller M, Stahel WA (2011) Sharpening wald-type inference in robust regression for small samples. *Comput Stat Data Anal* 55:2504–2515.
44. Cromwell JB, Labys WC, Terraza M (1994) *Univariate Tests for Time Series Models* (Sage Publications, Thousand Oaks, CA), Vol 99.