Kashyap and Verkroost *EPJ Data Science* (2021) 10:39
https://doi.org/10.1140/epjds/s13688-021-00294-7

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

Analysing global professional gender gaps using LinkedIn advertising data

Ridhi Kashyap[1,2,3*] and Florianne C.J. Verkroost[1,2]

*Correspondence:
ridhi.kashyap@nuffield.ox.ac.uk
[1]Department of Sociology, University of Oxford, 42-43 Park End Street, OX1 1JD, Oxford, United Kingdom
[2]Nuffield College, University of Oxford, New Road, OX1 1NF, Oxford, United Kingdom
Full list of author information is available at the end of the article

Abstract

Although women's participation in tertiary education and the labour force has expanded over the past decades, women continue to be underrepresented in technical and managerial occupations. We analyse if gender inequalities also manifest themselves in online populations of professionals by leveraging audience estimates from LinkedIn's advertisement platform to explore gender gaps among LinkedIn users across countries, ages, industries and seniorities. We further validate LinkedIn gender gaps against ground truth professional gender gap indicators derived from the International Labour Organization's (ILO) Statistical Database, and examine the feasibility and biases of predicting global professional gender gap indicators using gender gaps computed from LinkedIn's online population. We find that women are significantly underrepresented relative to men on LinkedIn in countries in Africa, the Middle East and South Asia, among older individuals, in Science, Technology, Engineering and Mathematics (STEM) fields and higher-level managerial positions. Furthermore, a simple, aggregate indicator of the female-to-male ratio of LinkedIn users, which we term the LinkedIn Gender Gap Index (GGI), shows strong positive correlations with ILO ground truth professional gender gaps. A parsimonious regression model using the LinkedIn GGI to predict ILO professional gender gaps enables us to expand country coverage of different ILO indicators, albeit with better performance for general professional gender gaps than managerial gender gaps. Nevertheless, predictions generated using the LinkedIn population show some distinctive biases. Notably, we find that in countries where there is greater gender inequality in internet access, LinkedIn data predict greater gender equality than the ground truth, indicating an overrepresentation of high status women online in these settings. Our work contributes to a growing literature seeking to harness the 'data revolution' for global sustainable development by evaluating the potential of a novel data source for filling gender data gaps and monitoring key indicators linked to women's economic empowerment.

Keywords: LinkedIn; Digital Demography; Gender; Sustainable Development Goals; Data for Development

1 Introduction

Over the past decades, with the expansion of women's participation and completion of tertiary education, women's ability to access professional occupations and managerial positions in labour markets has also increased [1, 2]. Yet, despite these improvements in

Springer

education, women continue to be underrepresented in technical and managerial occupations requiring specialised, tertiary-level education. Based on the most recent statistics available from the International Labour Organization (ILO), women comprise 44.8% of those in managerial, professional and technical occupations, compared to a male share of 55.2% [3]. Looking at senior positions, the gender gaps are even larger. In the majority of the 67 countries with data from 2009 to 2015, fewer than a third of senior- and middle-management positions were held by women [4]. These gender inequalities emerge from a complex interplay of factors, such as differences between men and women in human capital and fields of study, gender discrimination and negative stereotypes associated with women in professional roles, challenges in combining work and family life that disproportionately disadvantage women, limited access to social capital, networks and resources for professional advancement, and gender differences in values and interests [5–10].

Reducing gender inequalities in economic, social and political domains is essential for the attainment of global sustainable development. The promotion of gender equality features prominently in the United Nations' (UN) Sustainable Development Goals (SDGs), a key instrument in setting the agenda around global development, both as a standalone goal (Goal 5) as well as in relation to other goals (e.g. access to education) [11]. Target 5.5 within SDG 5 on gender equality emphasizes women's economic empowerment and recognises the importance of ensuring "women's full and effective participation and equal opportunities for leadership at all levels of decision-making in … economic … life" [12]. Measuring and understanding gender gaps in skilled, technical and managerial occupations is integral to monitoring progress on SDG 5.

This paper analyses global professional gender inequality by leveraging aggregated information about LinkedIn's user population, available through its advertising platform, as a type of 'digital census' of this online population. The rapid uptake of social media sites such as LinkedIn over the 21st century has generated large and diverse online user populations across different platforms, and these online populations have in turn generated new types of social data through the 'digital traces' they leave behind on these platforms. A growing body of work in the area of digital demography [13–16] has sought to understand the demographic characteristics of users on different social media platforms, analyse their demographic representativeness, and examine their potential to study gender inequality [17–22]. In a complementary vein, with the call for a 'Data Revolution' in the context of the post-2015 global development agenda [23, 24], a growing body of research has sought to examine the potential of non-traditional big data sources for measuring different indicators and social processes linked to the SDGs. This body of research has used diverse types of big data sources, including mobile call log data (e.g. [25, 26]), night-time satellite data (e.g. [27]), web and social media data (e.g. [18, 22, 28–30]), to assess their utility for providing cost-efficient, timely, and more granular coverage to complement traditional data sources such as surveys or censuses. These data sources offer added value in the context of low- and middle-income countries where traditional data sources are often lacking, incomplete or outdated [31]. In relation to SDG 5 on gender inequality, recent work has used social media advertising data from Facebook and Google to model indicators linked to internet and mobile gender gaps [18, 22]. Our study builds on this work by examining social media advertising data from LinkedIn for gender inequalities in the professional domain.

LinkedIn is the world's largest social networking platform targeted at professionals with a user base of over 700 million spanning over 200 countries. The site is used for job-seeking, recruiting, networking and marketing, and its mission is "to connect the world's professionals to make them more productive and successful."[1] As technical, professional, and managerial occupations in the world of work have become virtually mediated, social networking sites such as LinkedIn hold the potential to enable economic opportunity, provide new sources of professional information, and expand networks strategically for job opportunities and career advancement. Although research on the use of social media and more specifically on LinkedIn for economic opportunities and career advancement is still limited, emerging survey and qualitative evidence suggest that users do indeed experience some of these benefits [32–37].

The payoffs of easier access to professional information and networks through platforms such as LinkedIn have the potential to be larger for those populations who otherwise face greater barriers to access these resources, such as women and ethnic or racial minority groups [7]. Existing evidence from the US points to a greater role for the informational and network benefits of internet and social media technologies for groups who are disadvantaged in the labour market [38, 39]. Studies focusing on low-income countries find that digital technologies have larger impacts on women for outcomes linked to health, well-being and economic opportunity, as women conventionally face greater barriers in accessing information and have smaller social networks compared to men [40, 41]. For economic opportunities, networking behaviours can offer a useful strategy to break through the glass ceiling for women, although empirical evidence suggests that men are more likely to engage in some forms of networking, such as socialising after work [42], which especially disadvantage mothers with child-care responsibilities. In this context, online and more flexible forms of networking offered by platforms such as LinkedIn could help overcome some of the barriers experienced by women in the offline world. On the other hand, although online platforms promise to be open, flexible, and democratic spaces, existing literature also shows that women are often significantly underrepresented in online communities catering to more specialised or technical environments such as those of Wikipedia editors [43, 44], StackOverflow [45] or GitHub users [46]. While these gender gaps in part reflect women's underrepresentation in technical fields such as software engineering, they also reflect factors such as gender biases within these online communities that disadvantage women, cultural or algorithmic features that discourage female participation and result in faster dropout among female users, and behavioural differences in the use of these platforms.

Beyond country-specific studies from the US [21, 47] and UK [48], or analyses focused on smaller samples for specific occupations [17, 49, 50], little is known about gender differences in the LinkedIn user population in a global, comparative perspective, as well how these gender differences on LinkedIn's user population vary across age, industries and seniority. More broadly in the research on social network sites, studies of LinkedIn are considerably fewer than those of Twitter or Facebook [48]. Similarly, while the growing body of research in digital demography has often drawn on aggregate audience estimates from social media advertising platforms to analyse these online populations, most of these

[1] https://about.linkedin.com/?trk=homepage-basic_directory_aboutUrl

studies have used Facebook (e.g [30, 51, 52]) or more recently, Google advertising (AdWords) audience estimates [22]. In contrast, LinkedIn's advertising audience estimates have so far only been used to study professional gender gaps for a selection of information and communication technology (ICT) industries [17], and for a selection of cities in the United States [21]. Both these studies point to the potential value of these LinkedIn data for studying socio-demographic phenomena.

The first objective of this paper is to analyse gender gaps on LinkedIn's user population by computing different country-level LinkedIn Gender Gap Indices (GGIs) by age, industry and seniority to examine how gender gaps manifest themselves across different characteristics on this online population. Our second objective is to compare and validate the LinkedIn gender gaps against ground truth indicators of country-level professional gender gaps derived from nationally-representative labour force surveys, available via the International Labour Organization's (ILO) Statistical Database (ILOSTAT), to examine the feasibility and biases of predicting ground truth measures using LinkedIn's online population. The novel data from LinkedIn cover a large number of users across the world and show wider geographical coverage than the data from ILOSTAT, including better coverage in low- and middle-income countries where data on these indicators are often limited or lacking. Nevertheless, gender gaps on LinkedIn are those measured on an online social media population and the biases of this social media population are not properly understood. By comparing against ground truth data, our analysis enables us to address whether online gender gaps on LinkedIn broadly reflect professional gender inequalities in the labour force across countries, or whether gender gaps observed on LinkedIn reveal unique gender biases and selection effects, and if so, how these are patterned.

2 Data sources and methods

2.1 Dataset

For our study, we build a dataset of country-level indicators derived from different sources, including those from audience estimates obtained from LinkedIn's advertising platform, labour force indicators obtained from ILOSTAT, as well as other indicators linked to global development (e.g. Human Development Index), ICT penetration (e.g. internet penetration), and gender equality in different domains (e.g. in educational attainment or internet access). We describe these different indicators below.

2.1.1 LinkedIn ad audience estimates

We rely on a type of digital trace data called 'advertising audience estimates', which are available for all large social media platforms such as Facebook, Google, Twitter, LinkedIn, and others [31]. The basic idea behind these data sources is similar across all platforms, even though the types of estimates provided varies across them. For example, while Facebook's ad platform provides counts of monthly or daily active users, Google AdWords provides ad impression estimates (i.e. the number of times an ad would be seen) rather than an estimate of users [22].

Potential advertisers on online platforms can specify a desired audience for their ads based on targeting criteria, such as gender, age, geography, and other characteristics. This serves as a kind of real-time digital census over the user base of the considered platforms, providing aggregate answers to questions such as "How many female Facebook users aged 18 years of age live in Nigeria?" Ad audience estimates from Facebook and Google have

been leveraged previously to model digital gender inequality indicators linked to internet access and mobile access gender gaps, which are targets within SDG 5 on promoting gender equality [18, 22, 53]. These studies motivate the assessment of ad audience estimates from other platforms to model gender inequalities in different domains. This paper considers indicators within SDG 5 linked to women's economic empowerment.

In contrast to Facebook or Google, which are platforms of broader appeal, LinkedIn is the world's biggest social networking site targeted at professionals. The platform is used for job-seeking, recruiting, networking and marketing. LinkedIn allows users including workers, employers or recruiters, to create a profile based on their professional affiliation and connect to professional contacts within and outside their professional networks. Users turn to LinkedIn for different reasons, including professional self-promotion, accessing information about career opportunities, organizations and professional development, strategic networking, and communication with professional contacts [32, 33, 35, 37, 54–56]. Similar to other social media platforms, one of the key aspects in LinkedIn's business model is to make revenue through targeted advertising offered on the platform through its ad campaign manager.[2] Audiences can be targeted on the basis of geographic location, demographic criteria such as gender or age group, and job criteria such as company industry or job seniority. Before advertisements are launched and shown to specific target audiences, the advertiser is shown an approximate number of how many individuals match the targeted audience. For example, at the time of data collection there were approximately 1.3 million male directors in the United Kingdom on LinkedIn. This feature allows us to collect data from a sizable population, which are available in real-time and capture about 17% of all ∼4 billion Internet users [57].

To build our dataset, we collected aggregate counts on numbers of LinkedIn users by querying the ad campaign manager via its application programming interface (API). The data were collected over February-April 2019, which corresponds approximately to the year with the most recently available labour force surveys (2019, followed by 2018) in the ILOSTAT, from where we obtain our labour force ground truth indicators (see next section). Using these aggregate counts, we generate different gender gap indicators of the LinkedIn population at the country level for as many countries for which these estimates were available. Previously, LinkedIn advertising data have been used to examine variation in global gender gaps in the ICT sector [17] and professional gender gaps in 20 US cities [21]. Both these studies find gender gaps disfavouring women on LinkedIn, but also document significant variations in these gender gaps by different characteristics. For example, [21] find that gender gaps in US cities vary by age with more men relative to women at older ages, and considerably by industry with education being female-dominant and construction being male-dominant. [17] find that when looking only at users within the information and communication technology industry on LinkedIn, women are significantly underrepresented compared to men, but that even within ICT, variations in sub-fields exist with male-bias being stronger in hardware industries compared to software.

We compute gender gaps on LinkedIn in terms of a gender gap index (GGI), which is defined as

$$\text{LinkedIn GGI} = \frac{\text{Number of females on LinkedIn with characteristic}}{\text{Number of males on LinkedIn with characteristic}} \quad (1)$$

[2] https://business.linkedin.com/marketing-solutions/ad-targeting

Characteristics for which we collect data in this study are countries, age groups, company industries, job seniorities, job function and field of study, which are always disaggregated by gender. For example, the most general LinkedIn GGI for a given country would be the LinkedIn overall GGI, which is defined as the ratio of the number of females to males where the data obtained are by gender and country. The LinkedIn age 25–34 GGI, for example, would be defined as the ratio of the number of females aged 25–34 to males aged 25–34 in a given country. This definition of a gender gap as a female-to-male ratio is akin to the gender gaps as defined in other global indices such as the Global Gender Gap Report [58]. Our choice of this formulation of a gender gap indicator is also aligned with other work modelling digital gender gaps using social media advertising data [17, 18, 22].

We remove all observations for which the total audience count (men plus women) is less than 1000 to capture larger audience sizes. This is also similar to the approach applied in other studies using social media advertising data in which country observations with small audience counts are filtered out to ensure more stability in estimates and to avoid highly specialised features or categories [18, 22]. As a result of this filter, observations are generally removed for countries with small populations (e.g. Bouvet Island, Cocos (Keeling) Islands, Cook Island, Falkland Islands (Malvinas), Montserrat, Svalbard and Jan Mayen, Wallis and Futuna, Tokelau, Tuvalu, Saint Pierre and Miquelon, S. Georgia and S. Sandwich Islands, etc.) and categories (e.g. specific language and literature studies (such as Khmer/Cambodian, Uralic or Ukrainian)) that are less represented on LinkedIn. Furthermore, the application of this filter does not change the observed correlations between the LinkedIn overall GGI and the three ILO GGIs that we describe later, because in the overall data the filtering removes only six countries for which the ILO GGIs were not available in the first place (i.e. Bouvet Island, Cocos (Keeling) Island, Heard and McDonald Island, S. Georgia and S. Sandwich Islands, Svalbard and Jan Mayen and Tokelau).

For any given country, we generally find that the larger the number of categories within a characteristic (e.g. field of study or company industry), the higher the number of zero audience counts in specific categories within the characteristic. In other words, data sparsity is greater in more detailed or specific categories. This is particularly the case for field of study (301 categories) and company industry (147 categories), where there are a considerable number of categories with zero counts, and to a smaller extent for job function (26 categories), job seniority (10 categories) and age group (4 categories). Gender was available in three categories – male, female and unknown – and we dropped unknowns when computing GGIs. Like with gender, unknown values exist for all characteristics, and these are dropped when computing any GGI. At the time of our collection, aggregate counts with non-missing gender and country information were available covering a population of 460.18 million users; with non-missing gender, country and age group information aggregate counts were available covering 165.02 million users; and non-missing gender, country and information on any industry (before merging these into more aggregate International Standard Industrial Classification of All Economic Activities (ISIC) 1 or Science, Technology, Engineering and Mathematics (STEM)/non-STEM categories) were available for 368.83 million users. Descriptive statistics summarising audience counts by gender for different characteristics as well as number of countries available for each are provided in Table 1.

Compared with traditional data sources such as labour market statistics, the LinkedIn data offer a number of advantages. LinkedIn data cover a large number of users across

Table 1 Descriptive statistics of LinkedIn data by gender and further disaggregated categories. Note: Number of countries includes those after the application of the observation filter with audience counts of fewer than 1000 removed

Category	Audience counts (in millions)			Number of countries with sufficient data
	Total	Women	Men	
Gender	460.18	196.06	264.12	234
Age group				
18–24	43.61	19.61	23.99	187
25–34	59.60	26.75	32.85	201
35–54	47.68	19.00	28.68	194
55+	14.14	4.53	9.61	133
Company industry				
Non-STEM	273.71	122.50	151.21	190
STEM	97.22	30.59	66.63	164
Field of study				
Non-STEM	152.87	73.62	79.25	175
STEM	70.51	17.98	52.53	147
Job function				
Non-STEM	226.35	106.55	119.80	203
STEM	87.53	25.40	62.13	201
Job seniority				
Chief X Officer (CxO)	8.14	2.01	6.14	137
Director	20.20	7.15	13.05	177
Entry-level	124.92	55.49	69.43	218
Manager	28.00	9.66	18.34	184
Owner	16.36	5.23	11.13	153
Partner	2.48	0.69	1.79	87
Senior	92.01	41.33	50.68	212
Training	6.14	2.82	3.32	126
Unpaid	5.03	2.11	2.92	98
Vice President (VP)	8.52	2.38	6.14	140

a large number of countries, provide harmonised data across them (which makes cross-country comparison easier) and offer the benefit of low latency. However, similarly to other social media ad audience estimates, the LinkedIn data also suffer from a number of weaknesses, such as issues of non-representativeness, limited metadata to understand the data generating process, and the potential for algorithmic confounding [31, 59]. To better understand their strengths and limitations, we therefore compare and validate them against external indicators as described in the next section to better understand who we are capturing online on this platform.

2.1.2 International Labour Organization's Statistical Database (*ILOSTAT*)

To validate the LinkedIn gender gap measures, we compare them to three different ground truth measures from the ILOSTAT, namely the ILO professional GGI [3], the ILO total management GGI [60] and the ILO senior and middle management GGI [4]. The data available through ILOSTAT are derived from the most recently available country-specific labour force surveys for each country, which can vary between countries. The modal year for the latest available labour force surveys in the ILOSTAT database was 2019, followed by 2018. On average, high-income countries have more recent data coverage, whereas in low-income countries, the last available labour force survey is older. We choose these three measures because they are closely aligned with the highly skilled population on LinkedIn, and capture different dimensions of professional gender inequality. The three gender gaps

are defined as follows:

$$\text{ILO professional GGI} = \frac{\text{Number of females in level 3 or 4 skilled occupations (ILO)}}{\text{Number of males in level 3 or 4 skilled occupations (ILO)}}, \tag{2}$$

$$\text{ILO total management GGI} = \frac{\text{Female share in total management (ILO)}}{100 - (\text{Female share in total management (ILO)})}, \tag{3}$$

$$\text{ILO senior and middle management GGI} = \frac{\text{Female share in senior and middle management (ILO)}}{100 - (\text{Female share in senior and middle management (ILO)})}. \tag{4}$$

For the ILO professional GGI, we only consider the highest International Standard Classification of Occupation (ISCO) skill levels 3 and 4 occupations because LinkedIn is mainly targeted at those in skilled and managerial occupations. Occupations classified as ISCO skill level 3 require complex and practical tasks as well as technical, factual and procedural knowledge, usually acquired during a 1–3 year degree following secondary education [61]. Skill level 4 occupations require problem-solving, decision-making and research skills as well as theoretical and analytical knowledge usually acquired during a 3–6 year degree in higher education. Both skill levels 3 and 4 require adequate numeracy and literacy as well as intercommunication skills. Among the three ILO GGIs in terms of data availability, country coverage is best for the ILO professional GGI, followed by the ILO total management GGI, and the worst for the ILO senior/middle management GGI.

In a similar way as for LinkedIn data, ILO data can be queried based on characteristics or strata like age group and economic activity, of which some are available by gender. In parts of our analyses, we further match age-specific and industry-specific LinkedIn indicators to age-specific and industry-specific ILO indicators to examine correlations between them. Note that these industries are referred to as "economic activity" in ILOSTAT. To do so, the ILO age group of 15–25 has been matched to LinkedIn's category of 18–24; ILO categories 35–44 and 45–54 have been merged to match LinkedIn's 35–54 age group; and ILO 55–64 and 65+ age groups have been combined to match LinkedIn's 55+ category. The industry-level matching between LinkedIn and ILOSTAT data has been done manually on the basis of International Standard Industrial Classification [62] sections (ISIC level 1) using matching data provided by [63].

2.1.3 Other development, ICT, and gender inequality indicators

Our dataset also includes other indicators linked to economic or human development, ICT penetration, and gender gaps in educational and ICT domains derived from different sources such as the UN, World Bank and World Economic Forum. We use these indicators to examine the biases of our predictions of professional gender gaps derived from the LinkedIn measures, as we describe in more detail in the upcoming Methods section. For general measures of economic and human development we include GDP per capita [64] as well as the human development index (HDI) for females, males and both [65]. The HDI is a composite measure capturing education, economic and health dimensions of well-being for a country's population. For ICT penetration, we include levels of internet penetration

[66], as well as online model estimates of gender gaps in internet access from the Digital Gender Gaps (DGG) project [53].[3] The choice of these indicators is motivated by prior work that draws on web or social media data, and these help to inform our expectations about the nature of the bias we may expect in the predictions generated from LinkedIn data.

At low values of internet penetration and the human development index, we may expect the online population to be more selective and less representative of the entire population [67]. Internet penetration, however, has been shown to be gender-differentiated, and particularly in regions of South Asia and Sub-Saharan and Northern Africa, women have lower levels of internet access than men [18, 22]. Internet access gender gaps may have implications for our gender gap predictions generated using LinkedIn data, although the direction of this bias is theoretically ambiguous. On the one hand, the fact that fewer women are online relative to men may lead us to overestimate professional gender inequality using LinkedIn data, as women are less likely to be online and may consequently use internet-related technologies such as LinkedIn less. On the other hand, existing literature also suggests that online selectivity can work in counter-intuitive ways. For example, Magno and Weber showed that in countries with greater offline gender inequality (e.g. Pakistan, Egypt), the online status of women as measured on Google+ was higher [68]. The authors speculated that this observed selectivity was due to a 'Jackie Robinson Effect', akin to the scenario 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 [69]. Similarly, [70] found that the women with profiles on Wikipedia were likely to be more notable than men, suggesting an analogous type of glass ceiling effect where women had to pass a higher threshold to be captured on this online population. Whether similar types of selectivity also manifest themselves on LinkedIn remains to be assessed.

To further examine these links between online and offline gender selectivity, we include indicators linked to gender gaps in educational attainment, enrolment in secondary and tertiary education and the labour force from the Global Gender Gap Report [58], and gender gap indices in ICT and STEM education [71]. The bias observed in predictions from LinkedIn may reflect gender differences in the use and incentives to use the platform, which may depend on the levels of gender equality in the educational and economic domains. However, the direction of the bias is theoretically ambiguous. On the one hand, social media platforms promise to be open and inclusive spaces with the potential to enable content-sharing, expanded access to information and the opportunity to connect with wider networks for those who may lack access to these resources through conventional channels [33–35, 37]. For example, women with children experience greater time pressures due to the dual burden of managing work with childcare responsibilities [42, 72]. In this context, social media may provide more flexible models of networking than those that involve post-work hours socialising, which significantly disadvantage women with childcare responsibilities. The incentives and benefits linked to the use of digital technologies may consequently be larger for women, particularly those who may have limited opportunities to expand their networks in other ways, such as those in countries where their educational or economic attainment is poorer, and/or those in male-dominated fields [40, 41].

[3] www.digitalgendergaps.org

This line of reasoning suggests that in countries where women's relative economic or educational attainment is lower, predictions from LinkedIn data may lead us to overestimate gender equality in the actual labour force.

On the other hand, studies of technical and specialised online communities such as GitHub, StackOverflow or Wikipedia editors show that women continue to be underrepresented on these online communities in a way that reinforces offline gender gaps. More specifically, these gender inequalities may arise due to the experience of gender discrimination based on observable information on profiles [46], cultural norms or algorithmic features of the platform that discourage female participation [43, 45, 73] and reward more male-oriented behaviours [74, 75]. In contrast to platforms such as StackOverflow or GitHub that cater to software communities, little is known about gender inequalities in the use and incentives to use LinkedIn, which caters to a broader range of professional occupations. Nevertheless, it is plausible that an overrepresentation of male users or male senior managers in countries with greater offline gender inequalities may further disincentivise female users from joining or participating, particularly if networking behaviours exhibit homophily by gender, as recent experimental evidence suggests [76].

In addition to these broader socio-demographic country-level factors, the biases observed may reflect the extent to which LinkedIn is used as a platform. To assess if our biases differ based on the degree of LinkedIn penetration among those in the highly skilled (ISCO skill levels 3/4) labour force in a given country, we compute overall levels of and gender gaps in LinkedIn penetration, where LinkedIn penetration is defined as the ratio of the number of people on LinkedIn and the number of people in highly skilled jobs in the ILOSTAT [3]. For example, if LinkedIn is a widely used platform in a given country, we might expect less bias than in countries where LinkedIn penetration is low. Further, differences in LinkedIn penetration may vary by gender, which may reflect compositional differences between the male and female populations on LinkedIn, or differences in behaviors or incentives to use the platform. An example of a compositional characteristic is age structure. To illustrate, if the LinkedIn population in a country comprises disproportionately of younger users and has a younger age structure, the aggregate gender gap indicator may reflect these younger cohorts more. As younger cohorts are likely to be more gender egalitarian [1, 2, 77], a larger presence of younger cohorts in a particular country's LinkedIn population would lead us to overestimate professional gender equality in predictions using LinkedIn data.

2.2 Methods

Our analysis proceeds in three steps. First, we perform descriptive analyses to examine how gender gaps on LinkedIn manifest themselves across different parts of the world and across different characteristics, such as age groups, job seniorities, job functions, company industries and fields of study. Second, we examine correlations between LinkedIn GGIs and the three ground truth measures obtained from ILOSTAT, also by age group and company industry as well as other external development and gender inequality indicators. We then build regression models using LinkedIn GGI measures to predict the three ground truth ILO gender gap measures. Our modelling exercise starts with the most parsimonious, one-variable linear regression model using the LinkedIn overall GGI, which has the best country coverage. In each regression table, we report the total number of observations for which the dependent and independent variables are available in the full

sample (N_{obs}) as well as the total number of observations for which predictions can be made (N_{pred}) in the full sample (i.e. for which the independent variables are available) to highlight the gains in coverage, if any, of using the LinkedIn data source. Linear regression allows us to avoid overfitting and to have interpretable results. Sensitivity analyses have shown that binomial regressions on the proportion of women of total (women / (women + men)) do not result in improved performance compared to linear regressions on the GGI. While the regression coefficients for the model are based on analyses on the full sample, we also provide measures of out-of-sample prediction performance using five-fold cross-validation. We use $k = 5$ for a reasonable bias-variance trade-off, and perform five repeats to reduce the bias in the estimator, given that the folds in non-repeated cross-validation are dependent (because samples used for training in one fold are used for testing in another). The measures of performance we report are the means across the folds and repeats of the mean absolute error (MAE), the root mean squared error (RMSE), and the R^2. Note that this cross-validation (CV) R^2 is the square of the Pearson's correlation coefficient of the observed and predicted values (as implemented in the `R` package `caret`) [78].

As our dataset contains a variety of LinkedIn gender gap indicators computed across different characteristics for a given country, such as age, industry and seniority, we further assess if using a wider range of LinkedIn variables can help improve the predictive performance of our single-variable model. We expand the number of candidate predictors to include the four age group values available (i.e. 18–24, 25–34, 35–54 and 55+), as well as gender gaps across seniorities (ten unique values) and company industries (147 unique values reduced to 19 ISIC level 1 classes, as explained in Sect. 2.1.2), which along with the LinkedIn overall GGI result in a total of 34 candidate predictors. Due to considerable data sparsity (i.e. zero audience counts) within unique job functions and fields of study, and due to their poor availability across different countries, we omit these two characteristics from this analysis. We consider an indicator a candidate when it is available for at least 125 countries or at least the number of countries for which the dependent ILO GGI is available, whichever is larger. Note that this does not necessarily mean that the regression will be performed on a sample of at least 125 countries, because missingness patterns across a selection of variables may result in a smaller number of complete observations across these variables. After the application of these filters, we are left with 17 candidate variables for the ILO professional GGI, 19 for the ILO total management GGI, and 34 for the ILO senior/middle management GGI. We use lasso regression to fit these models [79], where variable selection is performed across the available candidate predictors. To find the optimal value of shrinkage parameter λ in the lasso, which in turn influences which variables are selected, we use five-fold cross-validation on the data to choose the value of λ that minimises the MSE. The model corresponding to this optimal value of λ is then reported. We report the adjusted R^2 and RMSE for the lasso regression for the full sample, as well as error metrics (cross-validated (CV) R^2, MAE and RMSE) computed using five-fold cross-validation.

The third step of our analyses focuses on analysing patterns of bias in the professional GGI predictions generated using the LinkedIn GGI. For this, we focus on analysing the residuals (difference between the observed ILO GGIs and the ILO GGIs predicted from the LinkedIn overall GGI, i.e. $y - \widehat{y}$) of the aforementioned parsimonious single-variable model as our outcome of interest, with the different development and gender gap indi-

cators described in Sect. 2.1.3 as possible predictors. As our goal here is to better understand which variables are useful for explaining the bias of our predictions, we use hybrid (mixed forward and backward selection) variable selection to find the variables that maximize the adjusted R^2. We choose the adjusted R^2 as selection criterion because we want to maximize the variation explained in order to characterise those contexts where we systematically under- or overpredict professional gender inequality using the LinkedIn data.

3 Results

3.1 Describing global gender gaps on LinkedIn

Figure 1 shows the LinkedIn overall GGI for all countries for which we are able to compute it. A value below one indicates females are underrepresented relative to males on LinkedIn, whereas values exceeding one indicate female overrepresentation relative to males, and one indicates gender parity. While LinkedIn GGI values indicating greater gender inequality disfavouring women (GGI about 0.2) exist in most parts of Africa, the Middle East and some parts of Asia (including India and China), the proportion of women is closer to 50%, corresponding to a GGI of at least one, in most parts of the Americas, Europe and Oceania. Some of the most gender egalitarian countries – where women even outnumber men on LinkedIn – include Latvia, Lithuania, Moldavia, Georgia, Myanmar, Vietnam and Bhutan.

To further examine gender gaps across different characteristics for which we are able to collect audience counts for different countries, Fig. 2 and Table 2 show the LinkedIn GGI for different characteristics. Note that for field of study, job function and company industry, the categories are manually divided into either STEM or non-STEM here using the STEM-Designated Degree Program List [80] to generate broader aggregations. For example, the age group 35–54 represents the ratio of the number of women and men in this age group as obtained from LinkedIn data by country, gender and age group. While Fig. 2 shows the medians of all boxes lie underneath the dashed line (parity cut-off at GGI = 1 where male and female representation is balanced), there is substantial varia-

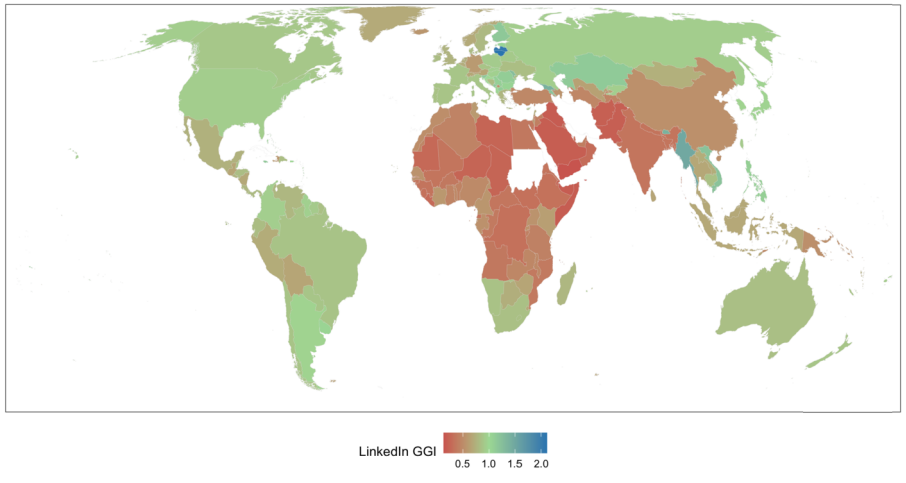

Figure 1 LinkedIn Gender Gap Index (GGI), computed as the female-to-male ratio of LinkedIn users in a country, across the world (New Equal Earth Projection)

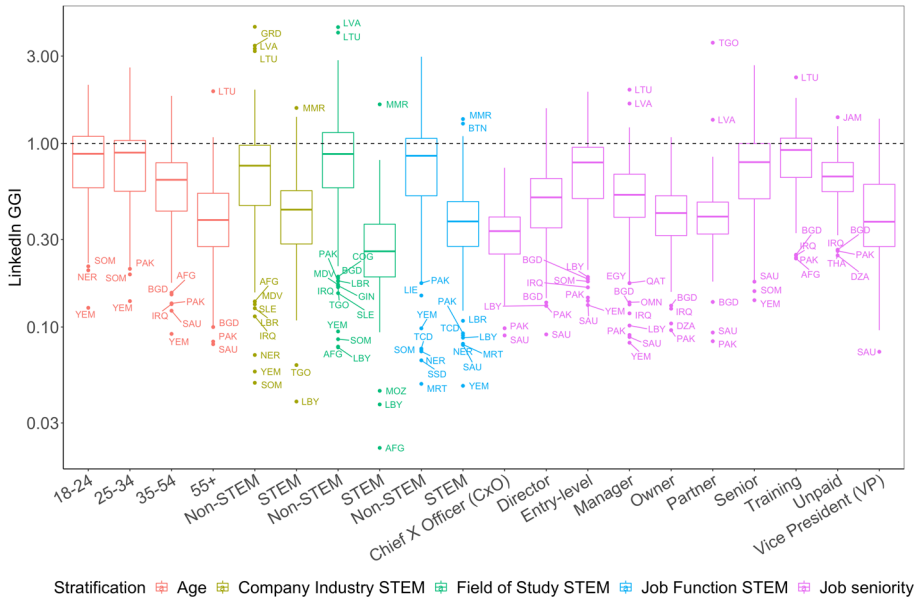

Figure 2 Box and Whisker plot for LinkedIn Gender Gap Index (GGI) by different characteristics. Note: *Y*-axis values are on a log (base 10) scale

tion in gender equality across these different characteristics. While there are almost as many women as men in younger ages, non-STEM industries, fields of studies and job functions, and low-level seniority jobs, gender inequality disfavouring women is highest among older people, in STEM fields, and among those with higher-level job seniority on LinkedIn.

Some recurring outliers that show LinkedIn GGIs with high values greater than one, i.e. countries where women outnumber men, include Latvia, Lithuania, Myanmar and Vietnam. Previous findings using these data to assess gender inequality in the information and communication technology (ICT) sector have also shown that these countries tend to be more gender egalitarian when it comes to LinkedIn users [17]. We further inspect these outliers against ILO ground truth data. Out of all 34 countries for which the overall LinkedIn GGI is larger than or equal to one, the ILO professional/technical GGI is also larger than or equal to one in 29 countries, missing in three countries and smaller than one in two countries. These observations support the notion that the "outliers" we observe in the LinkedIn data are not caused by measurement issues in those data, but rather actually represent countries where professional gender equality is relatively high.

Figure 3 shows the distribution across categories in age, field of study, company industry, job function and job seniority for women and men, aggregated across countries. Table 1 shows the corresponding audience counts across different categories by gender, as well as country coverage across different categories. Here, we observe that the female population on LinkedIn is relatively younger than the male population, that men on LinkedIn are more likely have or work in STEM fields of study, industries and job functions, and that women are more likely to hold entry- or senior-level jobs while men are more often Vice Presidents, Chief X Officers, directors, partners, owners and managers. We show two country-specific examples of one high-income country, the UK, and another lower-

Table 2 Summary statistics of LinkedIn Gender Gap Indices (GGI) by different characteristics (as displayed in Fig. 2)

Category	Min.	1st quartile (25% percentile)	Median (50% percentile)	3rd quartile (75% percentile)	Max.	Number of countries
Overall	0.13	0.51	0.73	0.89	2.12	234
Age group						
18–24	0.13	0.57	0.88	1.10	2.09	187
25–34	0.14	0.55	0.89	1.04	2.60	201
35–54	0.00	0.42	0.63	0.79	1.82	194
55+	0.00	0.27	0.38	0.53	1.93	133
Company industry						
Non-STEM	0.00	0.41	0.74	0.97	4.33	190
STEM	0.00	0.22	0.41	0.54	1.56	164
Field of study						
Non-STEM	0.08	0.57	0.88	1.15	4.31	175
STEM	0.00	0.16	0.24	0.36	1.64	147
Job function						
Non-STEM	0.00	0.50	0.86	1.06	2.98	203
STEM	0.00	0.25	0.37	0.48	1.36	201
Job seniority						
Chief X Officer (CxO)	0.00	0.24	0.33	0.40	0.74	137
Director	0.00	0.34	0.50	0.63	1.56	177
Entry-level	0.13	0.50	0.79	0.95	1.92	218
Manager	0.00	0.38	0.52	0.68	1.97	184
Owner	0.00	0.29	0.41	0.52	1.08	153
Partner	0.00	0.32	0.39	0.48	3.55	87
Senior	0.14	0.50	0.79	1.00	2.67	212
Training	0.00	0.65	0.92	1.07	2.30	126
Unpaid	0.25	0.55	0.66	0.79	1.39	98
Vice President (VP)	0.00	0.25	0.36	0.58	1.37	140

middle income country, India, both with large LinkedIn populations, of some of these patterns by age and ten disaggregated industries (classified by ISIC level 1 industries) in Table 7 in the Appendix. For these two countries, we also observe that women are younger than men on LinkedIn. The industry data show that professional, scientific and technical industries, along with information and communication technology industries are the two most represented industries on LinkedIn for these two countries. Compared to ILO data on these industries in these respective countries, we observe that LinkedIn

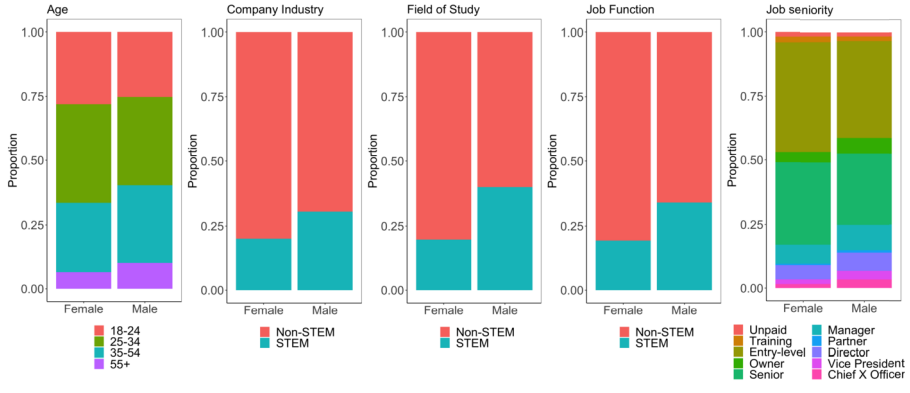

Figure 3 Gender composition of different characteristics among LinkedIn users

considerably overrepresents these industries for both men and women. Other industries such as manufacturing, education and health are also commonly represented on LinkedIn, but their share is comparatively underrepresented on LinkedIn in these countries. In the next section we examine correlations of gender gaps on LinkedIn against ILO ground truth data across different characteristics for all available countries to understand how broadly representative gender gaps across this online population and the labour force are.

3.2 Validating global gender gaps from LinkedIn

One of the main advantages of data on LinkedIn's user population is that they are generated as by-products of the platform's use, and can be queried in real-time. Labour force surveys, in contrast, are expensive to field and consequently not routinely available, especially in resource-constrained settings. Furthermore, the LinkedIn audience estimates are globally comparable and cover a large number of countries. While gender- and occupation-specific data from ILO only cover 190 countries, LinkedIn data cover 240 countries. More importantly, LinkedIn data have better coverage and recency for low and middle income countries. Nevertheless, LinkedIn data are from an online population of professionals, and it is unclear how the data observed on this platform compare with patterns in the labour force. Therefore, to explore the validity of LinkedIn data, we compare global gender gaps on LinkedIn to the three aforementioned ILO GGI measures from ILOSTAT. Figure 4 shows a scatter plot of the LinkedIn overall GGI and the ILO professional, total management and senior/middle management GGIs. The stars indicate the statistical significance of the correlations according to $^{*}p < 0.1$; $^{**}p < 0.05$; and $^{***}p < 0.01$. The correlation of the LinkedIn GGI with the ILO professional GGI is strongly positive at 0.71, and statistically significant at the 1% level. The correlation is even higher (0.81^{***}) when truncating GGI values of greater than one to one, as is the approach followed by commonly reported gender gap indices such as the Global Gender Gap report [58]. This suggests that the LinkedIn data may be better at capturing gender inequality in terms of women's disadvantage rather than capturing women's economic empowerment (where values exceed one). Despite these improvements in correlations, we choose not to truncate GGI values at one in our analyses as our objectives are not purely predictive but also to explore the

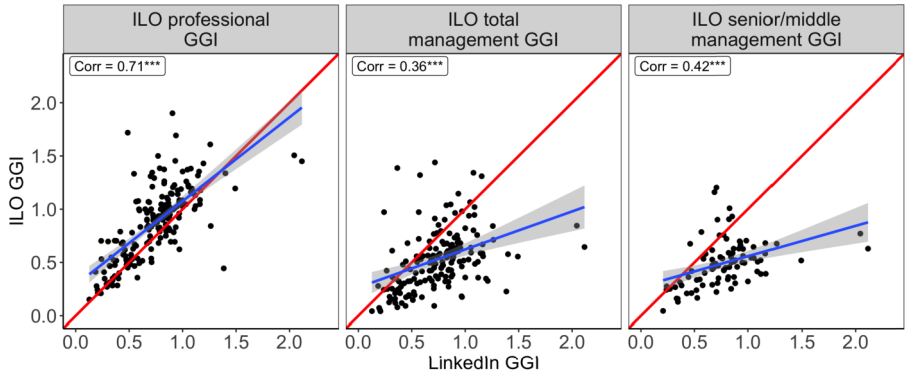

Figure 4 Scatter plot of LinkedIn overall GGI (*x*-axis) and ILO professional, total and senior/middle management GGI (*y*-axis)

indicator and its biases at different levels of gender inequality, including when it is high as well as low.

To further explore the correlations at different levels of professional gender inequality, we re-estimate these correlations removing the 34 aforementioned "outliers" from the data (e.g. Vietnam, Myannmar, Latvia) that display significantly high levels of professional gender equality both on LinkedIn and in the ILO data. Here too we find the observed correlations with the ILO GGIs are incrementally higher when these countries with very high levels of professional GGI values are removed than when containing these observations. That is, the correlation of the LinkedIn overall GGI with the ILO professional/technical GGI increases from 0.709 to 0.714; that with the ILO total management GGI increases from 0.416 to 0.526; and the correlation with the ILO senior/middle management GGI increases from 0.356 to 0.363. These observations again point to the idea that at higher levels of professional gender equality, LinkedIn data are broadly less representative of gender gaps in the labour force.

The red line in Fig. 4 is the $x = y$ equality line and the figures also report the Pearson's correlation coefficient (Corr) between the LinkedIn GGI and the three ILO GGIs. Looking at this, we see that generally, the ILO professional GGI values lie above the diagonal relative to the LinkedIn overall GGI, especially at lower levels of gender equality. In other words, there are relatively fewer women compared to men on LinkedIn than in the professional labour force. The correlation of the LinkedIn overall GGI with the total and senior/middle managerial ILO measures is 0.36 and 0.42 respectively (both are statistically significant at the 1% level), which is lower than with the ILO professional GGI but nevertheless reasonably high. In contrast to the ILO GGI indicators, the LinkedIn overall GGI is generally larger in value for a given country for both managerial ILO GGI measures and ILO senior/middle management GGI.

Figures 5 and 6 show the correlation between age-specific and industry-specific LinkedIn and ILO GGIs, respectively. The correlation between the age-specific LinkedIn and corresponding age-specific ILO measures is higher for older age groups, with almost no relationship among 18–24 year olds and a correlation of 0.54 among 55+ year olds. In other words, gender gaps in older ages appear to be generally more representative of the working population in professional occupations for older than younger age groups. We observe a greater variation in the younger age groups (18–24 and 25–34) in both the LinkedIn GGI and ILO professional GGI indicators. In contrast, there is much less variation across countries in professional gender equality among older age groups (men outnumber women in almost all countries), which might help explain the improved correlation in these age groups. Alternatively, LinkedIn users among younger age groups may also use LinkedIn in more varied ways, including for finding new opportunities and networking while still in education, and might therefore not be captured in the ILO professional GGI indicator which refers more specifically to individuals employed in these jobs.

The industries for which LinkedIn and ILO GGIs correlate strongly include education, finance, professional/technical/science, human health and social work, other services, real estate and public administration. Low correlations occur in agriculture, transportation and storage, mining and quarrying, wholesale and retail trade and manufacturing. These patterns generally suggest better correlations among more highly skilled sectors than lower skilled sectors, and more active use of LinkedIn in these higher skilled sec-

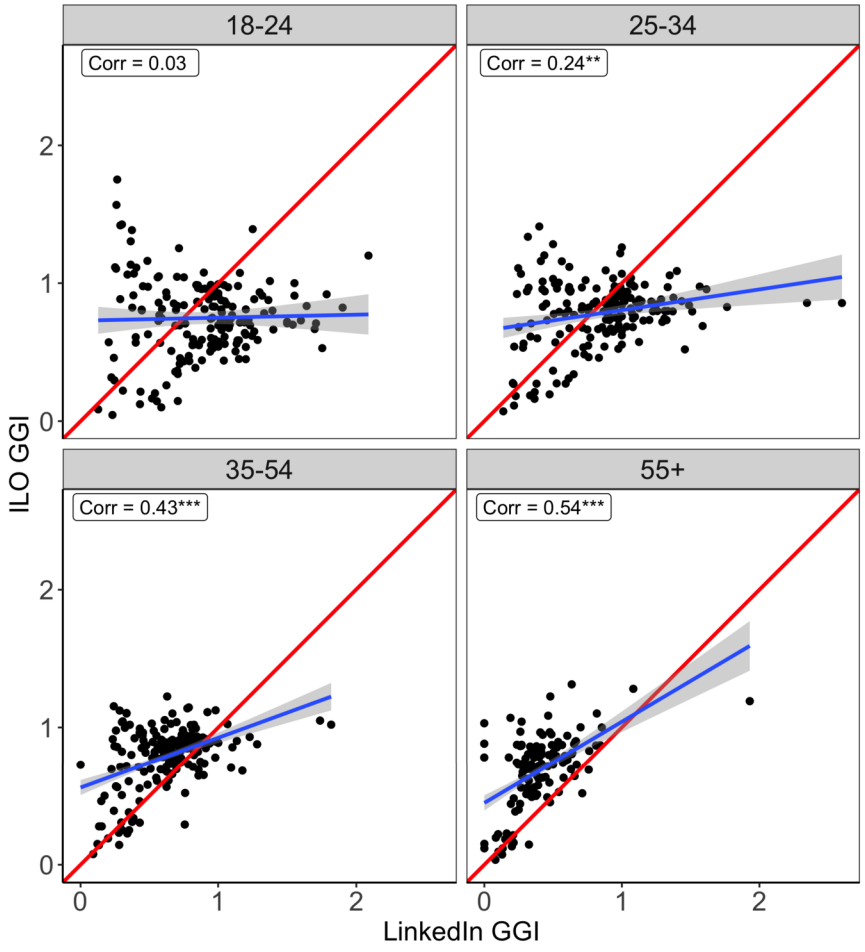

Figure 5 Scatter plot of age-specific LinkedIn GGI (*x*-axis) and ILO GGI (*y*-axis)

tors could underpin these patterns. One exception, however, is the information and communication industry, which is a highly skilled industry but for which the correlation between ILO and LinkedIn GGIs is relatively weak, albeit still positive and significant. This could be linked to greater variations in gender gaps across countries in this sector in both data sources [17], and/or gender differences in incentives to use LinkedIn across countries.

How do external measures of development and gender inequality compare to the correlation of the LinkedIn overall GGI with the ILO GGI measures? Table 3 shows correlations between the LinkedIn overall GGI, the three ILO GGIs and the other development and gender inequality indicators outlined in Sect. 2.1.3. The ILO professional GGI correlates most strongly with the LinkedIn overall GGI (0.71) – more strongly than other socioeconomic indicators in the dataset, such as the ILO senior and middle management GGI (0.66), and the GGIs in economic opportunity (0.57) and educational attainment (0.55) from the Global Gender Gap Report. These results suggest that this simple indicator from LinkedIn's online population has significant value as a predictor for the ILO GGI.

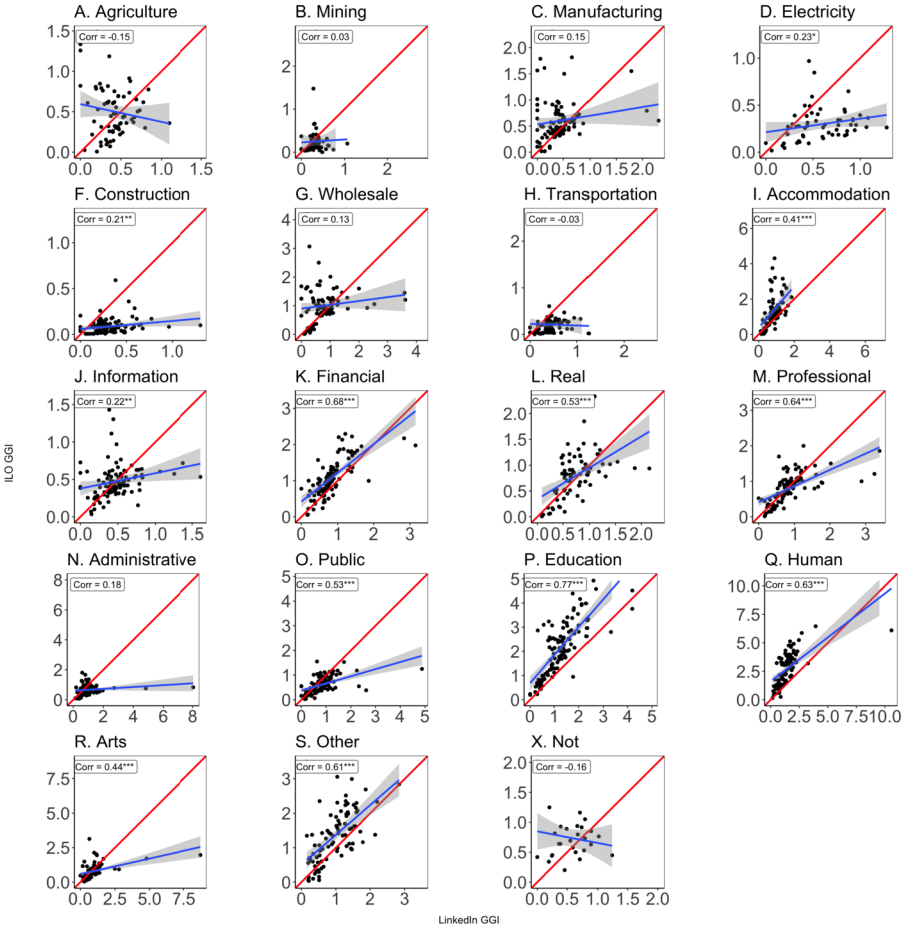

Figure 6 Scatter plot of industry-specific LinkedIn GGI (*x*-axis) and ILO GGI (*y*-axis). Note: Full industry names are: A. Agriculture; forestry and fishing; B. Mining and quarrying; C. Manufacturing; D. Electricity; gas, steam and air conditioning supply; "F. Construction; G. Wholesale and retail trade; repair of motor vehicles and motorcycles; H. Transportation and storage; I. Accommodation and food service activities; J. Information and communication; K. Financial and insurance activities; L. Real estate activities; M. Professional, scientific and technical activities; "N. Administrative and support service activities; O. Public administration and defence; compulsory social security; P. Education; Q. Human health and social work activities; R. Arts, entertainment and recreation; S. Other service activities; X. Not elsewhere classified

The ILO total management GGI correlates most strongly with the economic opportunity GGI (0.57), followed by similar correlations with the ILO professional GGI (0.49), ILO labour force participation GGI (0.37), LinkedIn overall GGI (0.36) and ILO senior and middle management GGI (0.35). The ILO senior and middle management GGI additionally correlates with the educational attainment GGI (0.47), LinkedIn overall GGI (0.42) and economic opportunity GGI (0.33). Interestingly, the GGIs in ICT and STEM education are generally negatively correlated with other measures of gender inequality and development. Although seemingly unexpected, these results are in line with the "STEM gender equality paradox" [81]. This paradox asserts that in low and middle income countries, women tend to study and work in STEM fields more often than in high income countries, in order to provide them with higher economic security. Overall, the ILO measures correlate strongly with the LinkedIn overall GGI. These findings indicate that our LinkedIn overall GGI indicator is a relatively good measure that is broadly representative

Table 3 Correlation table for LinkedIn, ILO and external development and gender inequality indicators

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1)	1.00***	0.71***	0.36***	0.42***	−0.23**	0.05	0.27***	0.53***	0.43***	0.25***
(2)		1.00***	0.49***	0.66***	−0.27***	−0.09	0.33***	0.55***	0.57***	0.17**
(3)			1.00***	0.35***	−0.23**	−0.24**	0.37***	0.19**	0.57***	−0.07
(4)				1.00***	0.05	0.18	0.11	0.47***	0.33***	0.13
(5)					1.00***	0.77***	−0.44***	−0.19*	−0.31***	−0.14
(6)						1.00***	−0.42***	−0.06	−0.27***	−0.05
(7)							1.00***	0.12	0.81***	0.13
(8)								1.00***	0.17**	0.42***
(9)									1.00***	0.14
(10)										1.00***

Note: *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.
Note: 1 = LinkedIn overall GGI; 2 = ILO professional GGI; 3 = ILO total management GGI; 4 = ILO senior and middle management GGI; 5 = UNESCO education ICT GGI; 6 = UNESCO education STEM GGI; 7 = ILO labour force participation GGI; 8 = GGG educational attainment GGI; 9 = GGG economic opportunity GGI; 10 = GDP per capita.

Table 4 Regression coefficients from the models predicting ILO GGIs from LinkedIn overall GGI for all countries (total) and countries with low (ILO GGI < median ILO GGI) and high (ILO GGI ≥ median ILO GGI) professional gender inequality

	Dependent variable:								
	ILO professional GGI			ILO total management GGI			ILO senior/middle management GGI		
	Total	Low	High	Total	Low	High	Total	Low	High
Intercept	0.29***	0.28***	0.93***	0.27***	0.16***	0.86***	0.27***	0.16***	0.73***
	(0.05)	(0.04)	(0.08)	(0.06)	(0.03)	(0.10)	(0.06)	(0.04)	(0.08)
LinkedIn overall GGI	0.79***	0.56***	0.24***	0.36***	0.23***	−0.12	0.29***	0.26***	−0.07
	(0.06)	(0.07)	(0.08)	(0.07)	(0.05)	(0.10)	(0.07)	(0.05)	(0.08)
N	185	92	93	167	82	85	89	44	45
N (pred)	234	92	93	234	82	85	234	44	45
R^2	0.50	0.45	0.09	0.13	0.24	0.01	0.17	0.36	0.02
5-fold CV									
CV R^2	0.52	0.50	0.14	0.20	0.31	0.07	0.24	0.37	0.11
MAE	0.17	0.12	0.14	0.20	0.10	0.20	0.15	0.08	0.13
RMSE	0.24	0.15	0.19	0.29	0.11	0.27	0.20	0.09	0.17

Note: *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

of professional gender inequality in the labour force, albeit clearly not one with a perfect correspondence with the ground truth gender gaps.

3.3 Predicting ILO GGIs with LinkedIn GGIs

Table 4 shows the regression coefficients for the models predicting the three ILO GGIs from the overall LinkedIn GGI for three different settings: all countries (total), countries where gender equality in the ILO GGIs is lower (ILO GGI < median ILO GGI) and countries where gender equality in ILO is higher (ILO GGI ≥ median ILO GGI). The coefficient on the LinkedIn overall GGI is positive and statistically significant for all ILO measures and settings, except for the high-equality settings for the two ILO managerial GGIs. The coefficient is close to one particularly in the total (0.79) setting for the ILO professional GGI, which our LinkedIn measure seems to predict best.

The best performance on the full sample is achieved on the ILO professional GGI (with an $R^2 = 0.5$) relative to the other ILO GGI measures linked to management and senior management. We note that data availability of the ground truth ILO indicators for model

fitting is also best for the ILO professional GGI indicator compared with the other two indicators. Looking further at the cross-validation fit metrics that illustrate the performance of these parsimonious models on unseen data in Table 4, we observe that the best performance in terms of RMSE, MAE and R^2 is similarly achieved on the ILO professional GGI, followed by the ILO senior/middle management GGI and the total management GGI. Further, we find that better predictive fit, both on the full-sample and cross-validation metrics, is achieved on the sample of countries in the low equality scenario relative to the high equality scenario. The models perform worse across the high equality scenarios across all three ILO GGI indicators. Consistent with the correlations reported earlier, these results further suggest that LinkedIn gender gaps are better at predicting professional gender gaps in settings where gender inequality disfavouring women is larger. In other words, when women are missing from the LinkedIn population in these low equality settings, this serves as a reasonable predictor that they are not in skilled professions in the labour force altogether. In contrast, when women are missing from the LinkedIn population in higher equality settings, this is less effective at predicting that they are not actually in the labour force in skilled professions, which may suggest more differentiated patterns of LinkedIn use in these settings. These patterns are consistent also with findings in [18] and [22] who found that Facebook and Google AdWords gender gap indicators were better able to predict internet access gender gaps in settings where internet access gender inequalities disfavouring women were larger.

The single-variable, parsimonious overall LinkedIn GGI performs well when it comes to predicting the ILO GGIs, albeit with better performance for the ILO professional GGI than for the other two managerial ILO GGIs. For all indicators, the ILO predictions generated using the LinkedIn GGI also come with improved country coverage, due to better data availability of the LinkedIn indicator, and this improvement in country coverage is greater for the two management-linked ILO GGIs. For the ILO professional GGI indicator, LinkedIn data enable predictions for 234 countries relative to 185 countries in the ground truth data, from 167 to 234 countries for the ILO total management GGI and for 89 to 234 countries for the ILO senior/middle management GGI. In terms of region, the biggest expansion in coverage occurs for countries in Africa.

The single-variable model relies on the overall LinkedIn population of users disaggregated by gender. We now consider the potential to improve predictive fit by considering LinkedIn gender gaps computed across other characteristics in our dataset. Table 5 shows the results from the lasso regression models where we consider gender gap indicators across age groups, job seniority and company industry as possible predictors.

The results in Table 5 show that for the ILO professional GGI, adding the GGI in senior jobs in addition to the LinkedIn overall GGI in the total scenario from Table 4 marginally increases the R^2 computed on the full sample from 0.50 to 0.55, as well as the CV R^2, and decreases the MAE and RMSE from the five-fold cross-validation. This however comes at the cost of decreasing the number of countries for which we can make predictions (234 in the simple model versus 209 with the more complex model). For the ILO total management GGI, gender gaps in director and manager job functions increases the R^2 from 0.13 to 0.14, but does not offer improved performance across the MAE, RMSE and R^2 computed using five-fold cross-validation.

The best increase in model performance occurs when predicting the ILO senior/middle management GGI from LinkedIn GGIs among VPs as well as those working in accom-

Table 5 Regression coefficients for model predicting ILO GGIs (variables selected using lasso regression and λmin parameter)

	Dependent variable:		
	ILO professional GGI	ILO total management GGI	ILO senior/middle management GGI
Intercept	0.39	0.31	0.21
LinkedIn overall GGI	0.01		
Job seniority			
Senior	0.59		
Director		0.32	
Manager		0.09	
Vice President (VP)			0.46
Company industry			
Accommodation and food service activities			0.19
Information and communication			−0.16
λ	0.02	0.02	0.02
N	159	143	54
N (obs)	174	144	65
N (pred)	209	177	123
RMSE	0.23	0.29	0.14
adj. R^2	0.55	0.14	0.42
5-fold CV			
CV R^2	0.56	0.16	0.35
MAE	0.17	0.20	0.13
RMSE	0.23	0.29	0.16

modation and food service activities and ICT, rather than from the LinkedIn overall GGI, resulting in an R^2 increase from 0.17 to 0.42, as well as improved performance in cross-validation MAE, RMSE and R^2. Overall, while these more complex models that rely on different characteristics appear to incrementally improve performance compared to the LinkedIn overall GGI model, it remains that our ability to use LinkedIn data to predict the total management and senior/middle management ILO GGIs is weaker than for the ILO professional GGI.

3.4 Explaining sources of bias in predicted global professional gender gaps from LinkedIn GGIs

Table 6 shows the variables selected when we model the residuals of our single-variable regression models (total scenario) reported in Table 4 using development and gender gap predictor variables. These predictors have been selected from a set of 14 aforementioned candidate variables using stepwise hybrid (combined forward and backward) selection, whereby candidates are entered and removed in a stepwise manner – based on the adjusted R^2 of the resulting models – until there aren't any predictors left for entering or removal.[4] Figs. 7, 8 and 9 show for all three ILO GGIs scatter plots of their actual (x-axis) and predicted (y-axis) values, whereby the points are coloured by the statistically significant (at $p < 0.05$) variables selected in each model as given by Table 6. Points that lie on the $x = y$ diagonal line indicate those countries for which the observed and predicted data correspond perfectly, whereas points above the diagonal are those where LinkedIn gen-

[4]Sensitivity analyses with forward and backward selection on different criteria (e.g. Akaike information criterion (AIC), p-values) have shown that this method resulted in the models for which the trade-off between the percentage of explained variation in the residuals and the number of observations in the sample is optimal.

Table 6 Regression coefficients for model predicting the residuals from the total scenario regressions in Table 4

	Dependent variable: Residuals from Total scenarios in Table 4		
	ILO professional GGI	ILO total management GGI	ILO senior/middle management GGI
Intercept	0.14 (0.19)	−0.12 (0.29)	−0.63** (0.31)
LinkedIn penetration GGI	−0.67*** (0.04)	−0.24*** (0.07)	−0.40*** (0.09)
GGG labor force GGI		0.15 (0.13)	−0.28** (0.14)
Proportion LinkedIn users aged 18–24		0.47 (0.29)	
Internet access GGI	0.84*** (0.14)	1.36*** (0.27)	1.00*** (0.25)
HDI (male)	−0.32** (0.15)	−0.51* (0.30)	
LinkedIn penetration	−0.09*** (0.02)		
GGG educational attainment GGI	−0.02 (0.27)		
GGG secondary education enrollment GGI		−0.63*** (0.23)	0.50* (0.27)
Internet penetration	0.08 (0.07)	−0.25** (0.13)	−0.30*** (0.10)
GGG tertiary education enrollment GGI		0.04 (0.04)	
N	129	120	70
Adj. R^2	0.76	0.35	0.47

Note: *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

der gaps overpredict professional gender equality and points below the diagonal are those where LinkedIn gender gaps underpredict professional gender equality.

The results in Table 6 show that significant predictors of the residuals in Table 4 common across all three ILO indicators are gender gaps in internet access and gender gaps in LinkedIn penetration. Other variables that emerge as significant predictors for at least one of the ILO indicators include overall levels of internet penetration, human development, and variables linked to gender gaps in education and labour force participation for a country. The proportion of variation explained is highest (76%) for the residuals for predicting the ILO professional GGI, and variables linked to LinkedIn penetration, gender gaps in internet access, as well as levels of human development (males) are statistically significant.

The proportion of variation explained for the residuals for the two managerial ILO indicators is less than the ILO professional GGI. For the ILO total management GGI, the selected variables are able to explain around 35% of the variation in the residuals. The selected, statistically significant variables include LinkedIn penetration GGI, internet access GGI, internet penetration, gender gaps in secondary educational enrollment and human development index (males). Gender gaps in internet access, LinkedIn penetration, in labour force participation, secondary education, as well as and levels of overall inter-

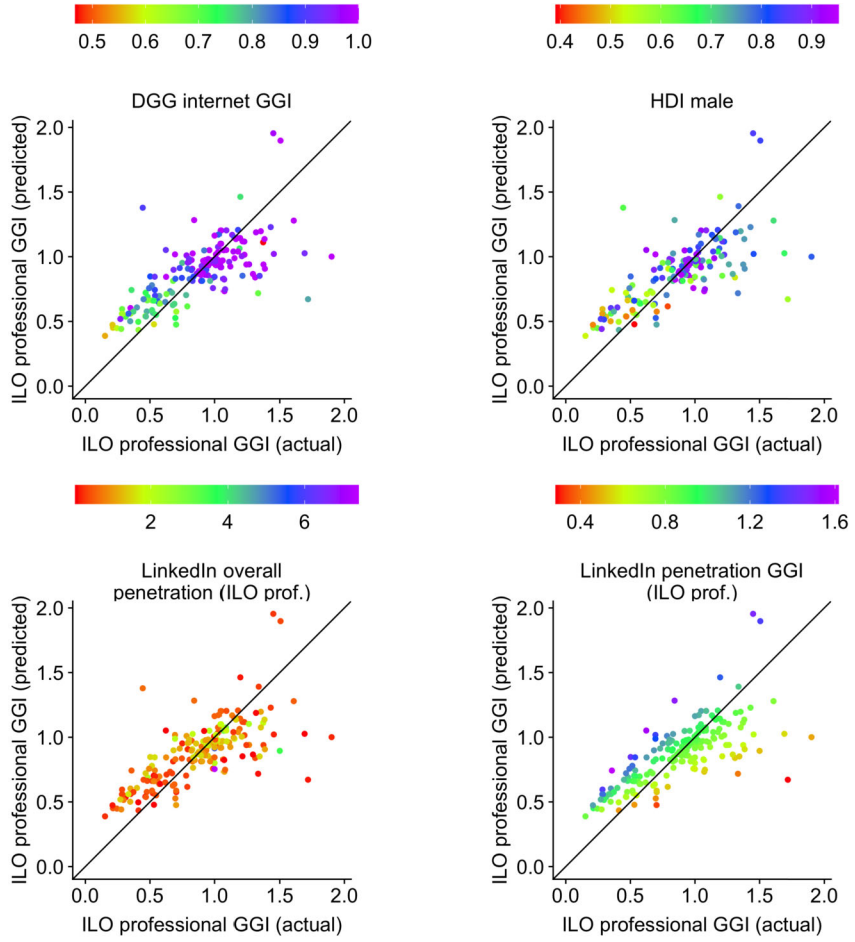

Figure 7 Scatter plot of ILO professional GGI (actual) (*x*-axis) and ILO professional GGI (predicted) (*y*-axis). Points in each panel are colour coded by values of the variable listed

net penetration explain about half of the variation (0.47) in the residuals in the ILO senior/middle management GGI.

The scatter plots in Figs. 7, 8[5] and 9 further help to visualize the direction and extent of the bias for the residual predictions. Across all three ILO indicators, the figures indicate that we tend to overpredict professional gender equality in countries with lower levels of gender equality in internet access (smaller internet GGI values). This pattern of bias indicates that in countries with greater gender inequality in internet access, higher status women are overrepresented online on LinkedIn, leading us to be more optimistic about professional gender equality in these settings than is actually the case in the labour force. This pattern of gender selectivity, where women are overrepresented on LinkedIn in countries where internet access gender gaps are larger, likely indicates the selective nature of the female online population in these settings. Across all three indicators, at higher values of the LinkedIn penetration GGI, we tend to overpredict gender equality than that observed in the ILO data. This suggests that the relatively higher penetration of LinkedIn among women relative to men in the labour force (i.e. when LinkedIn penetration GGI values

[5] We remove Togo, an outlier with a ILO total management GGI of 2.34, from this Figure so as to not skew the legend scale.

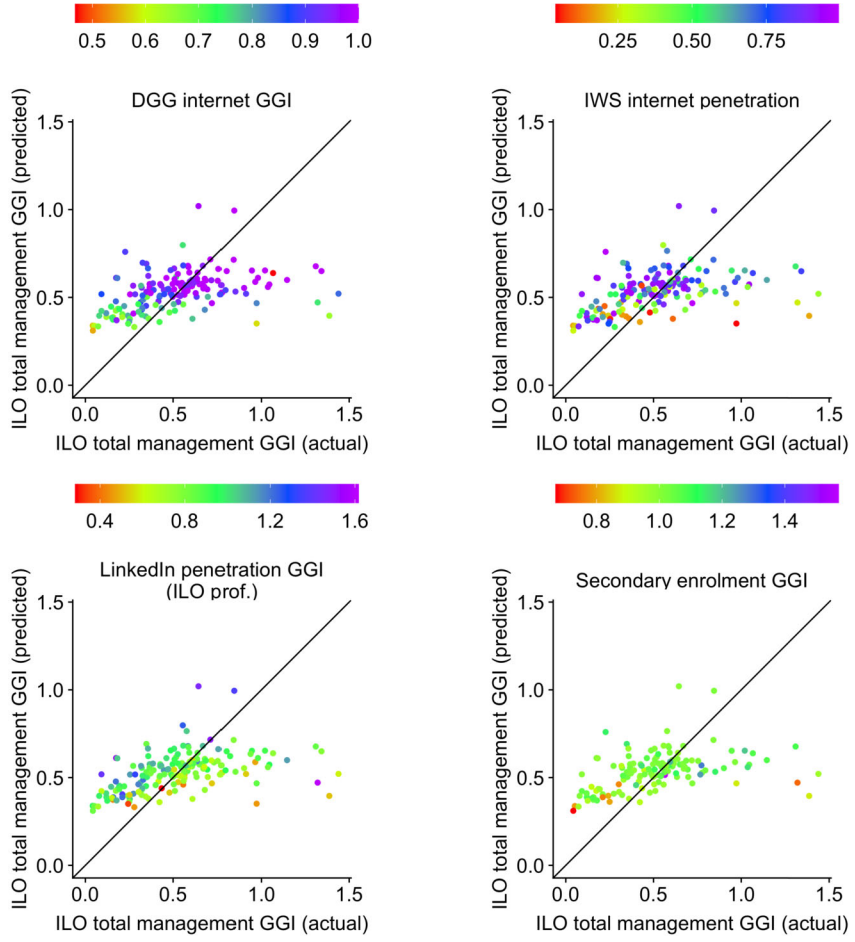

Figure 8 Scatter plot of ILO total management GGI (actual) (*x*-axis) and ILO total management GGI (predicted) *y*-axis). Points in each panel are colour coded by values of the variable listed

are greater than one) also results in our model overpredicting the gender gap in favour of greater gender equality. The predictions tend to be more accurate when LinkedIn penetration is gender balanced (close to 1).

For the two managerial gender gap indicators (Figs. 8 and 9) the model tends to underpredict gender equality at higher levels of the ground truth. For the ILO total management GGI, in addition to the already noted patterns of bias linked to internet access gender gaps and LinkedIn penetration, Fig. 8 also shows that in countries where secondary gender gaps are very low (<0.8), the model predicts greater gender equality in total management gender gaps than observed in the ground truth. We see a similar tendency of the model to overpredict gender equality in senior/middle management in Fig. 9 at very low values of the labour force GGI. In addition to the patterns of gender selectivity linked to internet access discussed above, the overrepresentation of women on LinkedIn in countries where gender gaps in economic and educational attainment show greater inequalities could reflect a stronger need for women in these contexts to make themselves more visible, and leverage opportunities offered by digital technologies such as LinkedIn to improve their economic prospects.

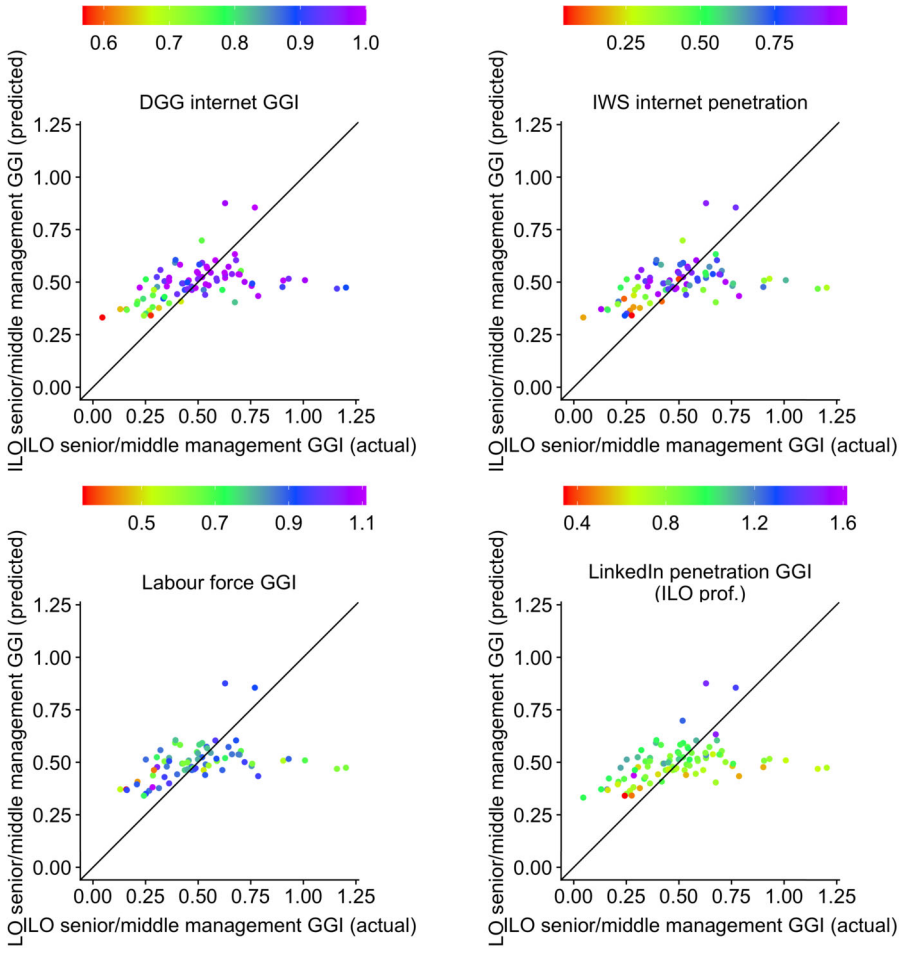

Figure 9 Scatter plot of ILO senior/middle management GGI (actual) (*x*-axis) and ILO senior/middle management GGI (predicted) (*y*-axis). Points in each panel are colour coded by values of the variable listed

Furthermore, Figs. 7, 8 and 9 show that our predictions are more accurate (points closer to $x = y$ line) at higher levels of internet penetration and/or LinkedIn penetration, and where internet access is gender balanced (close to 1). Essentially, when in a country internet use is more widespread and gender balanced in its population, and LinkedIn is more widely adopted among those in the work force, our predictions of professional gender equality are more accurate than in countries where fewer people are on the internet or on LinkedIn.

4 Discussion and conclusion

4.1 Summary

This paper leverages a novel source of digital trace data, LinkedIn's ad audience estimates available from its advertising platform, to analyse global professional gender inequality. To the best of our knowledge, this is the first global analysis of professional gender gaps on LinkedIn, alongside an assessment of the feasibility of using these data to model global gender gap indicators linked to women's economic empowerment in the context of the global sustainable development goals (SDGs). This study builds on other proof-of-concept

studies that have used the same data source to study gender gaps in the context of US cities [21], and gender gaps in the IT industry [17].

Our work contributes to a growing body of literature in digital demography, which examines the demographic characteristics and biases of online populations and the potential of digital trace data to measure and model socio-demographic processes. We document that, as with the offline world where women's participation in professional, technical and managerial occupations remains disadvantaged relative to men, the online world of LinkedIn also displays gender gaps, albeit with significant variations across countries, age groups, industries and seniorities. We find that gender inequality is particularly high in Africa, the Middle East and parts of Asia, while the Americas, Europe and Oceania are more gender egalitarian when it comes to having a LinkedIn profile. Additionally, gender inequality is larger among older individuals, those having studied or working in STEM sectors and industries, and among higher levels of job seniority on LinkedIn's population.

Our simple aggregate LinkedIn measure of gender equality, the LinkedIn Gender Gap Index (GGI), correlates positively and strongly with gender gap indicators from traditional labor force surveys available from the ILO, although these correlations are stronger for the general professional gender gap indicator than those linked to managerial or senior managerial gender gaps. These correlations indicate that LinkedIn gender gaps are more broadly representative of gender inequalities in technical and skilled populations in the labour force, while being less representative of managerial gender inequalities. We also find stronger correlations with ground truth measures for some industries (e.g. health, finance and insurance, education) over others (e.g. mining, agriculture, construction). The LinkedIn GGI also strongly correlates with other measures of gender equality in educational and professional domains, which suggests the potential for future work to explore the use of these data for modelling other outcomes linked to global sustainable development.

A parsimonious single-variable model using the LinkedIn GGI provides very good performance for predicting the ILO professional gender gap, and the model's ability to predict professional gender gaps is better in low professional gender equality contexts. This indicates that the absence of women from the online LinkedIn population in these settings serves as a good predictor that they are not in the professional labour force altogether. The model using LinkedIn data also expands country coverage of professional gender gaps beyond the ILO data. Although models that draw on LinkedIn gender gaps across other characteristics help improve predictive performance of the ILO senior/middle management gender gaps, these more complex models come with lower country coverage as more detailed features are not available for many countries.

Despite being a promising indicator, our work highlights interesting biases in professional gender gap predictions generated using the LinkedIn GGI. We find that our predictions are generally more accurate in countries with better and more gender balanced internet penetration, as well as LinkedIn penetration. A striking bias we find is that in countries with less gender egalitarian internet access, LinkedIn tends to predict greater gender equality than the ground truth. Similar patterns also emerge with the managerial indicators and gender gaps in educational and labour force indicators. These findings, which align with those in [68] for Google+, suggest that higher status women are disproportionately overrepresented online where connectivity gaps might be larger, and economic

barriers faced by women are greater. They also suggest the possibility that women in disadvantaged contexts may seek to signal their visibility more in online settings to overcome the barriers they face, or draw on digital technologies to improve access to resources or networks they cannot access through other channels. This is consistent with other findings from low-income countries that show that digital technologies have the potential to lower the costs of information, connectivity and networks, with the potential for bigger payoffs for those who face greater barriers such as women [40, 41]. Further work, both qualitative and quantitative, is needed to understand how gender differences across diverse contexts emerge in the use of the internet and social networking sites, and the implications of these gender gaps for economic and labour market outcomes.

4.2 Limitations

We acknowledge a number of limitations with our study. First, like all studies using social media advertising data, how user characteristics are determined to provide audience counts is not documented and historic estimates of past user counts are not available to track retrospective changes. Although LinkedIn states that this information is inferred from characteristics that individuals put on their profiles, how exactly this is done remains unknown. A further caveat here is that these data only allow us to assess whether women or men have a user account and do not say anything further about the intensity or types of use on LinkedIn, or duration of being a LinkedIn user. Research on gender-differential types of use and self-presentation on LinkedIn, while limited and based on small, specific samples from the US (e.g. a cohort of MBA graduates), suggests that women may provide less custom information on their profiles [49] and present emotions more while men present status [50]. It could be, for example, that the variation we capture across different industries, or seniorities, also reflect gender-differential patterns of disclosure of these characteristics rather than the presence of these groups on LinkedIn altogether. Moreover, little is known about the gender differences in the experiences of using this platform. Our work points to the presence of gender gaps on the platform, particularly among some groups such as senior managers, but the mechanisms for why these inequalities arise cannot be captured by our data. Further work is needed to understand these processes of disclosure of professional information and engagement on these platforms, as well as their implications for networking and economic opportunities.

LinkedIn data come from an online, social media population and our study has sought to understand how representative it is by comparing against ground truth indicators computed from labour force surveys, as well as analysing patterns of bias in predictions generated using LinkedIn data. Our work has shown that LinkedIn gender gaps are more predictive of general professional gender gaps indicators than those linked with managerial gender gaps. Particularly with the managerial gender gap indicators, more work is needed to understand how predictions using LinkedIn data sources could be improved. This may involve considering a wider range of features from within LinkedIn, although this may come at the risk of reduced country coverage as our work has shown. Other more promising approaches involving post-stratification [59] or the adoption of correction factors (e.g. [16, 18]) could serve as useful extensions. The variables we have shown to be useful in explaining the bias of our predictions could be incorporated into these models directly to help improve the performance of predictions of ground truth data.

4.3 Conclusions

Despite these limitations, our work provides evidence of how social phenomena in the offline world – in this case, professional gender inequalities in the labour force – also manifest themselves on the online world, yet with distinctive biases. It further illustrates the value of data generated from a novel online data source for monitoring policy-relevant social indicators within the global sustainable development goals, with the potential to fill important gender data gaps by complementing traditional data sources. The analysis we have presented provides one particular snapshot or cross-section, and our correlations show that online gender gaps capture cross-country variation in professional gender gap indicators in the ILO ground truth well. Further work is needed to examine if this extends to modelling changes over time through routinely collecting these data prospectively and comparing them against more ground truth indicators from labour force surveys as they become available. This kind of approach, in which more data from labour force surveys and ongoing data from LinkedIn are integrated, could be usefully applied for nowcasting professional gender gaps to provide higher frequency coverage than that provided by conventional labour force statistics, particularly in low income countries where ground truth data are often limited. Other applications might be to use these data to evaluate the impacts of specific policies or interventions on professional gender inequality and gender biases in online populations of professionals. Country-specific analyses drawing on more detailed characteristics on LinkedIn (e.g. skills) that are not often or easily captured in existing traditional data, or information for more local labour markets at more spatially granular levels would also be valuable extensions from the perspective of using these data for measuring policy-relevant indicators.

The gender gaps we document in this online, social media population have implications both for researchers as well as job recruiters and prospective employers. For researchers they indicate how social media populations such as LinkedIn display distinctive biases, particularly in contexts with greater internet access inequalities. This necessitates further work to better document and understand the nature of these biases, and emphasises the need to correct for these biases when studying or recruiting participants for research from such online populations. The presence of these biases also indicates the need for integrating different types of complementary data for a more inclusive understanding of social processes. For recruiters and prospective employers, our results suggest the use of LinkedIn could help to diversify hiring by gender in some contexts and cases – e.g. by recruiting younger or early career women who are more likely to be on the platform, in specific industry contexts, and in countries where women's educational or economic attainment is low, where we find that women are overrepresented relative to the labour force.

Appendix

Table 7 Proportions of LinkedIn audience counts by age and company industry, disaggregated by gender for India and the UK

Category	India				United Kingdom			
	Female		Male		Female		Male	
	LinkedIn	ILO	LinkedIn	ILO	LinkedIn	ILO	LinkedIn	ILO
Age group								
18–24	0.43	0.10	0.37	0.12	0.19	0.12	0.14	0.11
25–34	0.40	0.26	0.40	0.26	0.37	0.23	0.30	0.23
35–54	0.16	0.51	0.20	0.46	0.34	0.45	0.38	0.45
55+	0.01	0.13	0.03	0.15	0.09	0.20	0.18	0.21
Company industry (ISIC 1)								
A. Agriculture; forestry and fishing	0.00	0.57	0.00	0.40	0.00	0.01	0.00	0.01
B. Mining and quarrying	0.01	0.00	0.03	0.00	0.01	0.00	0.02	0.01
C. Manufacturing	0.08	0.13	0.14	0.12	0.05	0.05	0.10	0.13
D. Electricity; gas, steam and air conditioning supply	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.01
F. Construction	0.02	0.05	0.04	0.14	0.03	0.02	0.08	0.12
G. Wholesale and retail trade; repair of motor vehicles and motorcycles	0.05	0.05	0.06	0.12	0.11	0.12	0.08	0.13
H. Transportation and storage	0.02	0.00	0.03	0.06	0.02	0.02	0.04	0.08
I. Accommodation and food service activities	0.01	0.02	0.02	0.02	0.03	0.06	0.02	0.05
J. Information and communication	0.32	0.01	0.27	0.01	0.10	0.03	0.14	0.06
K. Financial and insurance activities	0.09	0.01	0.08	0.01	0.07	0.04	0.08	0.04
L. Real estate activities	0.01	0.00	0.02	0.00	0.02	0.01	0.02	0.01
M. Professional, scientific and technical activities	0.16	0.01	0.15	0.01	0.17	0.07	0.15	0.08
N. Administrative and support service activities	0.05	0.01	0.03	0.01	0.06	0.05	0.05	0.05
O. Public administration and defence; compulsory social security	0.01	0.01	0.01	0.02	0.05	0.07	0.03	0.06
P. Education	0.08	0.09	0.04	0.03	0.09	0.16	0.05	0.06
Q. Human health and social work activities	0.04	0.03	0.03	0.01	0.09	0.22	0.04	0.06
R. Arts, entertainment and recreation	0.03	0.00	0.03	0.00	0.08	0.03	0.06	0.03
S. Other service activities	0.01	0.02	0.01	0.02	0.04	0.04	0.02	0.02
X. Not elsewhere classified	0.00		0.00		0.00	0.00	0.00	0.00

Note that Table 7 shows the reduced versions of the ISIC level 1 industries; the full levels are as follows: A. Agriculture; forestry and fishing; B. Mining and quarrying; C. Manufacturing; D. Electricity; gas, steam and air conditioning supply; "F. Construction; G. Wholesale and retail trade; repair of motor vehicles and motorcycles; H. Transportation and storage; I. Accommodation and food service activities; J. Information and communication; K. Financial and insurance activities; L. Real estate activities; M. Professional, scientific and technical activities; "N. Administrative and support service activities; O. Public administration and defence; compulsory social security; P. Education; Q. Human health and social work activities; R. Arts, entertainment and recreation; S. Other service activities; X. Not elsewhere classified.

Acknowledgements
We thank Karri Haranko for helpful exchanges and for sharing code with us. The research for this paper was conducted as a part of the project 'The Digital Traces for the Gender Digital Divide' that received funding from Data2X, an initiative of the United Nations Foundation (Grant No. UNF-17-936). RK acknowledges support from the Leverhulme Trust through the Leverhulme Centre for Demographic Science. The funders had no role in the design and intellectual content of the study.

Abbreviations
GGI, Gender Gap Index; ILO, International Labour Organization; STEM, Science, Technology, Engineering and Mathematics; UN, United Nations; SDG, Sustainable Development Goal(s); ICT, Information and Communication Technology; ILOSTAT, International Labour Organization's Statistical Database; ISCO, International Standard Classification of Occupations; ISIC, International Standard Industrial Classification of All Economic Activities; HDI, Human Development Index; CV, Cross-validation.

Availability of data and materials
The dataset with compiled gender gap indicators and analysis scripts supporting the conclusions are available at https://github.com/fverkroost/epjds-professional-gender-gaps.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
RK conceptualised and designed the study. Both authors built the dataset. FV conducted the analyses in discussion with RK. Both authors interpreted the results and wrote the manuscript. Both authors reviewed and approved the final manuscript.

Author details
[1]Department of Sociology, University of Oxford, 42-43 Park End Street, OX1 1JD, Oxford, United Kingdom. [2]Nuffield College, University of Oxford, New Road, OX1 1NF, Oxford, United Kingdom. [3]Leverhulme Centre for Demographic Science, University of Oxford, 42-43 Park End Street, OX1 1JD, Oxford, United Kingdom.

Publisher's Note

Received: 16 March 2021 Accepted: 7 June 2021 Published online: 29 July 2021

References

1. Hannum E, Buchmann C (2005) Global educational expansion and socio-economic development: an assessment of findings from the social sciences. World Dev 33(3):333–354
2. Kc S, Barakat B, Goujon A, Skirbekk V, Sanderson WC, Lutz W (2010) Projection of populations by level of educational attainment, age, and sex for 120 countries for 2005–2050. Demogr Res 22(15):383–472
3. International Labour Organization (2020) Employment by sex and occupation (thousands) – Annual [Data set]. https://www.ilo.org/shinyapps/bulkexplorer5/?lang=en&segment=indicator&id=SDG_0552_OCU_RT_A
4. International Labour Organization (2020) Female share of employment in senior and middle management (%) [Data set]. https://www.ilo.org/shinyapps/bulkexplorer5/?lang=en&segment=indicator&id=SDG_0552_OCU_RT_A
5. Brass DJ (1985) Men's and women's networks: a study of interaction patterns and influence in an organization. Acad Manag J 28(2):327–343
6. Ragins BR, Sundstrom E (1989) Gender and power in organizations: a longitudinal perspective. Psychol Bull 105(1):51–88
7. Ibarra H (1993) Personal networks of women and minorities in management: a conceptual framework. Acad Manag Rev 18(1):56–87
8. Anker R (1998) Gender and jobs: sex segregation of occupations in the world. International Labour Organization, Geneva
9. Metz I, Tharenou P (2001) Women's career advancement: the relative contribution of human and social capital. Group Organ Manage 26(3):312–342
10. Hoobler JM, Lemmon G, Wayne SJ (2011) Women's underrepresentation in upper management: new insights on a persistent problem. Organ Dyn 40(3):151–156
11. Nations U (2015) Transforming our world: the 2030 agenda for sustainable development. Division for Sustainable Development Goals, New York
12. United Nations (2015) Sustainable Development Goal 5: achieve gender equality and empower all women and girls. https://sdg-tracker.org/gender-equality
13. Weber I, State B (2017) Digital demography. In: Proceedings of the 26th International Conference on World Wide Web Companion, pp 935–939
14. Cesare N, Lee H, McCormick T, Spiro E, Zagheni E (2018) Promises and pitfalls of using digital traces for demographic research. Demography 55(5):1979–1999
15. Alburez-Gutierrez D, Zagheni E, Aref S, Gil-Clavel S, Grow A, Negraia DV (2019) Demography in the digital era: new data sources for population research
16. Ribeiro FN, Benevenuto F, Zagheni E (2020) How biased is the population of Facebook users? Comparing the demographics of Facebook users with census data to generate correction factors. In: 12th ACM Conference on Web Science, pp 325–334
17. Verkroost FCJ, Kashyap R, Garimella VRK, Weber I, Zagheni E (2020) Tracking global gender gaps in information technology using online data. In: McDonald M (ed) Digital skills insights 2020. International Telecommunication Union, Geneva, pp 81–93. https://academy.itu.int/sites/default/files/media2/file/Digital%20S%kills%20Insights%202020.pdf
18. Fatehkia M, Kashyap R, Weber I (2018) Using Facebook ad data to track the global digital gender gap. World Dev 107:189–209
19. Garcia D, Kassa YM, Cuevas A, Cebrian M, Moro E, Rahwan I, Cuevas R (2018) Analyzing gender inequality through large-scale Facebook advertising data. Proc Natl Acad Sci 115(27):6958–6963
20. Mejova Y, Gandhi HR, Rafaliya TJ, Sitapara MR, Kashyap R, Weber I (2018) Measuring subnational digital gender inequality in India through gender gaps in Facebook use. In: Proceedings of the 1st ACM SIGCAS conference on computing and sustainable societies. ACM, New York, p 43
21. Haranko K, Zagheni E, Garimella K, Weber I (2018) Professional gender gaps across US cities. In: Twelfth International AAAI Conference on Web and Social Media
22. Kashyap R, Fatehkia M, Tamime RA, Weber I (2020) Monitoring global digital gender inequality using the online populations of Facebook and Google. Demogr Res 43:779–816

23. United Nations Secretary-General's Independent Expert Advisory Group on a Data Revolution for Sustainable Development (2014) A world that counts–mobilising the data revolution for sustainable development. Technical report
24. International Union for the Scientific Study of Population (2015) The IUSSP on a data revolution for development. Popul Dev Rev 41(1):172–177. https://doi.org/10.1111/j.1728-4457.2015.00041.x
25. Blumenstock J, Cadamuro G, On R (2015) Predicting poverty and wealth from mobile phone metadata. Science 350(6264):1073–1076
26. Mao H, Shuai X, Ahn Y-Y, Bollen J (2015) Quantifying socio-economic indicators in developing countries from mobile phone communication data: applications to Côte d'Ivoire. EPJ Data Sci 4(1):15
27. Elvidge CD, Sutton PC, Ghosh T, Tuttle BT, Baugh KE, Bhaduri B, Bright E (2009) A global poverty map derived from satellite data. Comput Geosci 35(8):1652–1660
28. Reis BY, Brownstein JS (2010) Measuring the impact of health policies using Internet search patterns: the case of abortion. BMC Public Health 10(1):1–5
29. Resce G, Maynard D (2018) What matters most to people around the world? Retrieving better life index priorities on Twitter. Technol Forecast Soc Change 137:61–75
30. Fatehkia M, Tingzon I, Orden A, Sy S, Sekara V, Garcia-Herranz M, Weber I (2020) Mapping socioeconomic indicators using social media advertising data. EPJ Data Sci 9(1):22
31. Weber I, Kashyap R, Zagheni E (2018) Using advertising audience estimates to improve global development statistics. ITU J ICT Discov 1(2)
32. Brouer RL, Stefanone MA, Badawy RL, Egnoto MJ, Seitz S (2015) Losing control of company information in the recruitment process: the impact of linkedin on organizational attraction. In: 2015 48th Hawaii international conference on system sciences. IEEE Comput. Soc., Los Alamitos, pp 1879–1888
33. Utz S (2016) Is linkedin making you more successful? The informational benefits derived from public social media. New Media Soc 18(11):2685–2702
34. Sharone O (2017) Linkedin or linkedout? How social networking sites are reshaping the labor market. In: Emerging conceptions of work, management and the labor market (research in the sociology of work, vol 30, pp 1–31
35. Cho V, Lam W (2020) The power of LinkedIn: how LinkedIn enables professionals to leave their organizations for professional advancement. Internet Research
36. Garg R, Telang R (2018) To be or not to be linked: online social networks and job search by unemployed workforce. Manag Sci 64(8):3926–3941
37. Davis J, Wolff H-G, Forret ML, Sullivan SE (2020) Networking via LinkedIn: an examination of usage and career benefits. J Vocat Behav 118:103396
38. Kuhn P, Mansour H (2014) Is Internet job search still ineffective? Econ J 124(581):1213–1233
39. Karaoglu G, Hargittai E, Nguyen MH (2021) Inequality in online job searching in the age of social media. Inf Commun Soc, 1–19
40. Suri T, Jack W (2016) The long-run poverty and gender impacts of mobile money. Science 354(6317):1288–1292
41. Rotondi V, Kashyap R, Pesando LM, Spinelli S, Billari FC (2020) Leveraging mobile phones to attain sustainable development. Proc Natl Acad Sci 117(24):13413–13420
42. Forret ML, Dougherty TW (2001) Correlates of networking behavior for managerial and professional employees. Group Organ Manage 26(3):283–311
43. Lam STK, Uduwage A, Dong Z, Sen S, Musicant DR, Terveen L, Riedl J (2011) Wp: clubhouse? An exploration of Wikipedia's gender imbalance. In: Proceedings of the 7th international symposium on wikis and open collaboration, pp 1–10
44. Hill BM, Shaw A (2013) The Wikipedia gender gap revisited: characterizing survey response bias with propensity score estimation. PLoS ONE 8(6):65782
45. Vasilescu B, Capiluppi A, Serebrenik A (2012) Gender, representation and online participation: a quantitative study of stackoverflow. In: 2012 international conference on social informatics. IEEE Comput. Soc., Los Alamitos, pp 332–338
46. Terrell J, Kofink A, Middleton J, Rainear C, Murphy-Hill ER, Parnin C (2016) Gender bias in open source: pull request acceptance of women versus men. PeerJ PrePrints 4:1733
47. Hargittai E (2015) Is bigger always better? Potential biases of big data derived from social network sites. Ann Am Acad Polit Soc Sci 659(1):63–76
48. Blank G, Lutz C (2017) Representativeness of social media in Great Britain: investigating Facebook, Linkedin, Twitter, Pinterest, Google+, and Instagram. Am Behav Sci 61(7):741–756
49. Are there gender differences in professional self-promotion? An empirical case study of Linkedin profiles among recent MBA. graduates
50. Tifferet S, Vilnai-Yavetz I (2018) Self-presentation in LinkedIn portraits: common features, gender, and occupational differences. Comput Hum Behav 80:33–48
51. Araujo M, Mejova Y, Weber I, Benevenuto F (2017) Using Facebook ads audiences for global lifestyle disease surveillance: promises and limitations. In: Proceedings of the 2017 ACM on Web Science Conference. WebSci '17. ACM, New York, pp 253–257. https://doi.org/10.1145/3091478.3091513
52. Zagheni E, Weber I, Gummadi K (2017) Leveraging Facebook's advertising platform to monitor stocks of migrants. Popul Dev Rev 43(4):721–734. https://doi.org/10.1111/padr.12102
53. Kashyap R, Weber I, Fatehkia M, Knowles I (2018) Digital Gender Gaps: measuring digital gender inequalities in real-time. www.digitalgendergaps.org/data
54. Papacharissi Z (2009) The virtual geographies of social networks: a comparative analysis of Facebook, linkedin and asmallworld. New Media Soc 11(1–2):199–220
55. Van Dijck J (2013) 'You have one identity': performing the self on Facebook and linkedin. Media Cult Soc 35(2):199–215
56. Baruffaldi SH, Di Maio G, Landoni P (2017) Determinants of phd holders' use of social networking sites: an analysis based on linkedin. Res Policy 46(4):740–750
57. International Telecommunication Union (2019) Measuring digital development: facts and figures. www.itu.int/en/ITU-D/Statistics/Documents/facts/FactsFigures2019.p

58. World Economic Forum (2019) Global gender gap report 2020. http://www3.weforum.org/docs/WEF_GGGR_2020.pdf
59. Salganik MJ (2017) Bit by bit: social research in the digital age. Princeton University Press, Princeton
60. International Labour Organization (2020) SDG indicator 5.5.2 – Female share of employment in managerial positions (%) | Annual [Data set]. https://www.ilo.org/shinyapps/bulkexplorer5/?lang=en&segment=indicator&id=SDG_0552_OCU_RT_A
61. International Labour Office (2012). International Standard Classification of Occupations 2008 (ISCO-08): structure, group definitions and correspondence tables. https://www.ilo.org/wcmsp5/groups/public/—dgreports/—dcomm/—publ/documents/publication/wcms_172572.pdf
62. United Nations Department of Economic and Social Affairs (2008) International Standard Industrial Classification of all economic activities (ISIC), Rev. 4. https://unstats.un.org/unsd/publication/seriesM/seriesm_4rev4e.pdf
63. World Bank (2019) Group & LinkedIn Corporation: LinkedIn industry to ISIC Rev. 4 industry mapping reference. [Data set]. https://development-data-hub-s3-public.s3.amazonaws.com/ddhfiles/144635/linkedin_to_isic_rev_4_industry_mapping_0.csv
64. World Bank Group (2020) GDP per capita, PPP (current international dollars) [Data set]. http://api.worldbank.org/v2/en/indicator/NY.GDP.PCAP.PP.CD?downloadformat=csv
65. United Nations Development Programme (2017) Human Development Data 2017. http://www.hdr.undp.org/en/data
66. Miniwatts Marketing Group (2018) Internet World Stats: usage and population statistics. https://www.internetworldstats.com/stats.htm
67. 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. WebSci '12. ACM, New York, pp 348–351. https://doi.org/10.1145/2380718.2380764
68. Magno G, Weber I (2014) International gender differences and gaps in online social networks. In: International conference on social informatics. Springer, Berlin, pp 121–138
69. Anzia SF, Berry CR (2011) The Jackie (and Jill) Robinson effect: why do congresswomen outperform congressmen? Am J Polit Sci 55(3):478–493
70. Wagner C, Graells-Garrido E, Garcia D, Menczer F (2016) Women through the glass ceiling: gender asymmetries in Wikipedia. EPJ Data Sci 5(1):5
71. United Nations Educational, Scientific and Cultural Organization (2019) Education: percentage of female graduates by field of study [Data set]. http://data.uis.unesco.org/index.aspx?queryid=164
72. Campbell KE (1988) Gender differences in job-related networks. Work Occup 15(2):179–200
73. Stephens M (2013) Gender and the geoweb: divisions in the production of user-generated cartographic information. GeoJournal 78(6):981–996
74. May A, Wachs J, Hannák A (2019) Gender differences in participation and reward on stack overflow. Empir Softw Eng 24(4):1997–2019
75. Vedres B, Vasarhelyi O (2019) Gendered behavior as a disadvantage in open source software development. EPJ Data Sci 8(1):25
76. Mengel F (2020) Gender differences in networking. Econ J 130(630):1842–1873
77. Percheski C (2008) Opting out? Cohort differences in professional women's employment rates from 1960 to 2005. Am Sociol Rev 73(3):497–517
78. Kuhn M et al (2008) Building predictive models in R using the caret package. J Stat Softw 28(5):1–26
79. Tibshirani R (2011) Regression shrinkage and selection via the lasso: a retrospective. J R Stat Soc, Ser B, Stat Methodol 73(3):273–282
80. US Immigration and Customs Enforcement (2012) STEM-Designated degree program list. https://www.ice.gov/sites/default/files/documents/Document/2014/stem-list.pdf
81. Stoet G, Geary DC (2018) The gender-equality paradox in science, technology, engineering, and mathematics education. Psychol Sci 29(4):581–593