The UK Online Gender Audit 2018: A comprehensive audit of gender within the UK’s online environment

Ana-Maria Huluba*, Jason Kingdon, Iain McLaren

Glass, London, United Kingdom

*Corresponding author.
E-mail address: anamaria.huluba@glass.ai (A.-M. Huluba).

Abstract

Gender inequality has exploded as a recent issue within mainstream media across US and UK cultural commentary. High-profile scandals of sexual harassment and gender pay differences have focused attention on the on-going disparity between sexes and political status. This paper presents a novel experiment in the application of so-called “big data” to analyse gender inequality. Using Artificial Intelligence (AI) techniques in the form of Natural Language Processing, a web crawler is used to audit the whole.uk online domain, and to measure the United Kingdom’s (UK’s) online economic presence for gender representation in terms of: prominence, job roles, and leadership within and across economic sectors. The procedure scans over 200 million web pages, and harvests 157,032 organisations and over 2.3 million people. The results reveal material bias (60%+) towards the representation of men over the majority of economic sectors, and across representation of power and status within job roles and professional titles. The experiment highlights not only new levels of gender bias but also the use of the Internet as a valuable source of plentiful data for social and economic analysis.

Keywords: Information science, Industry
1. Introduction

Gender equality is one of the 17 Sustainable Development Goals (SDGs) adopted by the United Nations in the 2030 Agenda for Sustainable Development. Gender biases appear to exist within all societies and the UK labour market is no exception (ONS, 2013; GOV.UK, 2018). The numbers and repeated patterns of bias suggest that some form of gender separation is either: voluntarily, tacitly, culturally or politically imposed (British Council, 2016; Egon Zehnder, 2016; Francoeur et al., 2006). In the age of the omnipresent web, the existence and prevalence of such biases may be more concerning as biases and stereotyping could entrench and promote such positions (Otterbacher et al., 2017; Fast et al., 2016). This paper aims to present a unique experiment in the application of so-called “big data” to the current high-profile issue of gender inequality within the UK. Using Artificial Intelligence (AI) techniques in the form of Natural Language Processing, an AI is used to systematically “read” the whole of the UK’s .uk online domain, so as to gather information relating to gender. The AI examines primary source material in the form of over 10 million websites, encompassing over 200 million individual web pages. It examines the number of men and women found, their job roles and titles, their economic sector, and their inferred power, or leadership, status.

The results are stark: gender representation online shows widespread biases. They reveal men and women seeking separate career paths and taking different jobs; highlight men “in control”, occupying the vast majority of leadership positions, and women embodying support and facilitation functions. It is shown that even in female-dominated sectors, men disproportionally bias the leadership positions.

Across a population of over 2.3 million people mentioned on the UK’s organisation websites it was found that: 92% of chairpersons are male, 82% of all CEOs are male, and 71% of directors are male. In contrast, 96% of legal secretaries are female, 94% of receptionists are female, and 78% of assistants are female. Only 5% of the time will you find a head of finance as a woman, slightly more likely than finding a man as senior executive assistant, a modest 3%. The BBC and the NHS were examined individually. It was found that BBC performers are over 70% biased towards men, as are directors and presenters. However, the female composer does even worse, with only 12% of the 138 musical composers featured on the BBC website.

The web data capture a unique snapshot of an “unselfconscious” web. Organisations have no legal obligations to present staff with balance, and have no expectation of being held to account for their choices, and so a website appears very much in the manner of its owners choosing. The results underscore the size of challenge to the equality movement.

This paper also highlights the emergence of web data as a new and important resource within social analysis. It is show that the harvested data tallies with official
data studies from the UK government (ONS, 2013), the WEF gender gap report (WEF, 2017), as well as regular industry surveys based on gender representation within UK businesses. The numbers presented here are in line with these traditional manual surveys - but have the advantages of being; i) materially larger in scope, and ii) more economic to attain. The results suggests that AI techniques may from a new and important role in the mining of data sources so as to conduct detailed forms of social and economic analysis.

The paper is also relevant in the context of ongoing research in the area of web data and such data’s ability to be used for statistical inference. For example, concerns over selection bias in other research contexts have been raised (Zagheni and Weber, 2015). In this study, however, bias is the subject matter being studied. It is the intention of the study to show what level of bias is prevalent online in respect of gender. It is the purpose to measure the level of male/female representation within UK-facing organisation’s websites. In this sense, the results are a further contribution to this important discussion. Finally, the work also supports other research efforts that make the case for using such data as a surrogate for official statistics (Dass et al., 2015; Askitas and Zimmermann, 2015). The findings presented here, show almost exact matching of official statistics, where such data exists, but with the advantage of greatly increased scale and, also the ability to see much deeper into the gender inequality issue, specifically around economic sectors and job roles. To the authors’ knowledge, this is the first such study presenting this depth of analysis and may also provide the potential for new measures and new means for tracking gender bias against policy initiatives.

This paper is organised as follows: Section 2 discusses advances in reading and interpreting web data, and the recent research in identifying gender disparities on the web. Section 3 describes the techniques used in this study to gather data, and infer gender. Section 4 presents the results. Section 5 discusses the findings and presents the limitations of this study. Finally, Section 6 concludes the paper and presents directions for future work in this area.

2. Related work

2.1. Reading web data

Web data is emerging as a discipline in its own right and concerns the mining of data entities from the so called world wide web (that is a subset of the Internet consisting of the pages that can be accessed by a Web browser — see www.techopedia.com). Entities are harvested from web pages and then aggregated to form a new body of data so extracted. In this sense, web data is aggregated, or restructured, from multiple primary data sources (the web documents, or pages, from the World Wide Web). A recent survey by Ferrar et al in 2014 (Ferrara et al., 2014;
Baumgartner et al., 2005) provides an overview and taxonomy of methods for web data extraction and discusses the range of uses that such data is applied to. The authors underscore this data’s unique potential (based on availability and quantity) to offer enhanced scope for analysis and deep examination of complex social and economic phenomena. They also show how such data has a variety of uses, from business, aimed to drive competitive advantage, through social science, decision support, and public policy making (Zagheni and Weber, 2015; Askitas and Zimmermann, 2015).

In this context, the automatic extraction of data from the web in order to support these needs has growing interest (Dong et al., 2014). One potential for this approach is solving some of the most demanding issues of survey data in that it is faster and cheaper to access, and requires limited human effort and so increases the scale and depth of analysis possible (Ferrara et al., 2014) and even introduces the ability to build time series.

Systems that store facts so harvested are also being created. These are referred to as knowledge bases, and typically retain information relating entities, such as people, places or organisations, together in structured data sets, schemas, ontologies and graphs (Paulheim, 2017).

For example, knowledge bases such as Freebase, YAGO, DeepDive or Knowledge Vault (Bollacker et al., 2008; Suchanek et al., 2007; Niu, 2012; Dong et al., 2014) use Natural Language Understanding techniques along with probabilistic measures to automatically derive knowledge or facts from the web.

However, the automatic reading and classification of web data in order to build such systems possesses some potential pitfalls too (Benfield and Szlemko, 2006). Data’s messiness and its structure have multiple challenges when it comes to extracting relevant facts and information. The main question in this area is how this data can be extracted, shaped into a semantically useful structures, and integrated into a system that allows further exploration and analysis (Baumgartner et al., 2005). This requires several trade-offs between automation and accuracy as well as challenges in the content structure, variability and privacy (Ferrara et al., 2014; Zaveri et al., 2016; Paulheim, 2017).

2.2. Gender disparities on the web

The analysis of gender disparities has a strong background, with studies examining the differences in career choices, corporate governance or management teams (Francoeur et al., 2006). Gender balances play a crucial role in the development of the society as an equal gender distribution across teams in organisations and sectors has been identified as significantly contributing to the organisational performance (Hoogendoorn et al., 2013).
The use of web data to better understand and describe gender disparities has come under academic attention largely due to the omnipresence of the web and the web’s potential for reflecting and influencing perceptions about gender roles in society. Disparities seem to occur everywhere: from workplace and job roles (Goldin, 2014) to images and word embeddings. For example, men are over-represented in online fiction stories (Fast et al., 2016) and also in news images (Jia et al., 2015). In contrast, web searches for images of occupations (e.g., nurse, investment banker), show that females are often portrayed and perceived as relatively less powerful (Kay et al., 2015) and are associated with “warm” character traits (Otterbacher et al., 2017). Reasons for such biases can be industry and skills related (Haranko et al., 2018).

Having a measure of gender differences plays a critical role in informing policymakers so that mitigation steps and polices can be taken, and in turn measured for impact and effectiveness. Official statistics, small-sample surveys, experimental tests, all contribute to a better understanding of this broader issue. However, the literature is currently lacking large-scale experiments carried at organisational level and that use open web data.

3. Methodology

3.1. Data collection

The focus of the study is on UK organisations that have a website. The sample is represented by public and private organisations that target a UK audience or have adopted the .uk internet domain address. An internet domain (e.g. www.bbc.co.uk) allows alphanumeric identification of any given website (Mockapetris, 1983). In an organisation context, a domain may correspond to a company’s trading name, one of its brand names or “represent a generic class of goods, services or interests” (Smith, 2007, p. 159). Countries can have their own domain namespace too. The UK’s domain name space is divided into several second-level domains: .co.uk, .ltd.uk, .plc.uk, .ac.uk, .gov.uk, .nhs.uk. and others.

The data characterising UK organisations was captured by an intelligent web crawler using Natural Language Processing techniques. The crawler reads and considered data from some 200 million individual web pages. Websites were harvested if:

- They were written in English;
- Had a UK postal address mentioned within their pages;
- Had some depth of representation for the organisation in question, that is, some description of the organisation that the AI could recognize;
- Had people (either mentioned through team pages, biographies, or roles or descriptions).
The aim was to seek organisation websites, and so the following were rejected for analysis: holding pages, low content pages, social media sites, blogs and services sites, such as shops, and search engines.

The human equivalent of the internet crawl would be to systematically browse websites addresses that end with a .uk domain, and look for “about-us” links (and equivalent), teams pages, and general site information pages, and then determine a certain level of quality regarding material so found.

Organisation name matching, and individual people detection, were based on natural language analysis, using sentence structures and syntax (see Sections 3.3 and 3.4).

3.2. Accuracy

Accuracy rates for the AI were validated by manual checking of over 5,000 individual random samples from each of the entity types captured. For the four main entities in this study, this consisted of over 20,000 accuracy trials. Accuracy as a percentage is simply how many times a human agreed with the AI that the entity had been correctly assigned. All data is available through Glass.ai by request. For the main entities this system scored the levels of accuracy given in Table 1. Each organisation discovered was also classified into one of the 108 economic sectors (see Supplementary Table for the full list).

3.3. Sectors

Sector classification is non-trivial, and an area of ongoing research as organisations perform cross-sector activities (Jones, 2013). To avoid the errors that this classification brings, for the purpose of this paper the LinkedIn taxonomy is used as one form of sector allocation as against many other options.

The LinkedIn site is popular social media site for business professionals. It is self-selecting in terms of how individuals choose to classify their own organisation sector. It has been used to analyse professional gender gaps or gender differences in a particular industry (Haranko et al., 2018; Yan and Ge, 2016). The LinkedIn sector classification represents at least one proposal of sector membership, and is

Table 1. Entity detection accuracy.

| Entity            | Accuracy |
|-------------------|----------|
| Company description | 95%      |
| Person            | 97%      |
| Biography         | 92%      |
| Job title         | 94%      |
beyond the scope of this study to examine possible categorisation errors, and issues raised by categorisation process.

3.4. People and gender inference

Once a qualified website was detected by the AI, the next phase was the collection of people data. On websites, people may be listed as employees, external members, or mentioned in news items. All these data were retrieved and analysed. There is no assertion about whether these people “really” are employed by, or indeed work with the organisation in question, instead, this paper is interested in the virtual “real-estate” allocated across gender lines.

A person was identified by a range of criteria, including the part-of-speech they appeared in, and how they were subsequently referenced in supporting text. Several heuristics were at work: firstly, the person’s title was considered an absolute gender identifier. Co-references by gender pronouns in supporting text provided additional evidence (Strazny, 2005). Finally, the gender was predicted based on an individual’s first name. The accuracy of the first name inference method was tested against the title and supporting pronoun as shown in Table 2. It can be noted that, the first name and title, and title and surname were considered to give 100% gender accuracy. There may be anxieties at this level of “certainty” in concluding gender type based on title and name. Is “Mr Henry Ford” really a man? This is a male representation, regardless of whether Mr Henry Ford, in the context of a website, even exists. In this sense, the same gender analysis could be applied in literary fiction, and the same assertions as to the representational biases offered could be made. Having cautioned this, the scale of data managed in this study suggest these inferences are accurate and correspond to the real-world situation. Great care was taken to filter organisation websites that for the most part present data in simple literal forms.

3.5. Dataset

Using the above criteria, from 10 million websites, the process obtained a total of 157,032 organisations websites with a qualified depth of data with a LinkedIn classification. On their websites more than 2.6 million people were mentioned, of which 87.92% had a determinable gender (Table 3).

### Table 2. Gender inference accuracy.

| Heuristic                  | Accuracy |
|----------------------------|----------|
| First Name                 | 97%      |
| First Name + Pronoun       | 98%      |
| First Name + Title – Pronoun | 100%    |
| Title                      | 100%     |
3.5.1. Internet domain distribution

The .co.uk is the dominant sub-domain level for the .uk domains, with over 90% of the organisations having this web address ending (Table 4). It is followed by .org.uk (4.47%) and nearly two times less schools .sch.uk. The domains and websites so gathered represents the “UK” universe of discourse in this study.

3.5.2. Economic sectors

Commercial organisations were separated into 108 economic sectors according to their LinkedIn profile. If no LinkedIn profile existed, the data was excluded. The top sectors discovered online are shown in Table 5.

Online presence is different from registered business, as not all UK organisations have a website (ONS, 2017a,b,c). Differences in the way sectors choose to represent themselves online is yet another area of future research. In this context, mapping or measuring online economic activity and comparing it, or modelling the interconnections with the real world, are related research concerns (Nathan and Rosso, 2015; Huws, 2014). For now, these aspects are noted, and imply that aggregated statistics will have this emphasis embedded within them. The ten lowest sectors are given in Table 6.

Table 3. The .uk domain gender audit data set.

| Data Set from .uk domains          | Data Sample |
|------------------------------------|-------------|
| Web sites                          | 157,032     |
| People                             | 2,657,286   |
| Gender assigned                    | 2,336,486   |
| Economic Sectors (LinkedIn)        | 108         |

Table 4. Domains distribution.

| Domain    | Organisations | %   |
|-----------|---------------|-----|
| .co.uk    | 141,651       | 90.21 |
| .org.uk   | 7,012         | 4.47 |
| .sch.uk   | 3,577         | 2.28 |
| .ac.uk    | 1,245         | 0.67 |
| .gov.uk   | 723           | 0.46 |
| .nhs.uk   | 589           | 0.38 |
| .ltd.uk   | 293           | 0.19 |
| .me.uk    | 107           | 0.07 |
| .police.uk| 36            | 0.02 |
| other     | 1,799         | 1.25 |
Table 5. Highest counts commercial sectors found in .uk domains.

| Sector                        | Organisations | %  |
|-------------------------------|---------------|----|
| Construction                  | 10,242        | 7.23|
| Hospitals and Medical Practices| 6,777         | 4.78|
| Law Practice and Services     | 5,701         | 4.02|
| Information Technology and Services | 5,443    | 3.84|
| Marketing and Advertising     | 5,296         | 3.74|
| Staffing and Recruiting       | 5,121         | 3.62|
| Real Estate and Property Management | 4,279    | 3.02|
| Design                        | 3,953         | 2.79|
| Internet                      | 3,856         | 2.72|
| General Financial Services    | 3,697         | 2.61|

Table 6. Lowest counts of sectors found online.

| Sector                              | Organisations | %  |
|-------------------------------------|---------------|----|
| Glass, Ceramics and Concrete        | 147           | 0.10|
| Computer Games                      | 137           | 0.10|
| Paper and Forest Products           | 120           | 0.08|
| Animation                           | 102           | 0.07|
| Gambling and Casinos                | 96            | 0.07|
| Semiconductors and Electronic Systems | 66          | 0.05|
| Investment Banking and Advisory     | 60            | 0.04|
| R&D and Scientific                  | 12            | 0.01|
| Libraries                           | 11            | 0.01|
| Tobacco                             | 10            | 0.01|

4. Results

Across all .uk domains, the headline representation of gender exactly tallies with the ONS\(^1\) official statistics for the gender split in the workplace [ONS September–November 2017](Table 7). Over such a large survey this seems very satisfying that the two estimates should so firmly agree, and suggests that online data can be an exciting new source of accurate economic and social analysis data.

4.1. Subdomain gender splits

If the headline picture seems to present a known outcome, then one layer down a mass gender segregation starts to emerge. Splitting the .uk domains into its subdomains gives the distribution from Table 8.

\(^{1}\) Office for National Statistics.
Standouts are schools at 74.47% women, the NHS 62.81% women, and the police at 66.2% male. Government at 40% female seems low, and .org, not-for-profit, and academia are the most balanced, as compared to the ONS gender split. The commercial sectors, in the form of .co.uk and ltd.uk show budding male bias. This was the next target for deeper analysis.

4.2. Gender splits according to economic sectors

Across all 108 sectors the minimum number of people found was for Tobacco (118 people), while the maximum (372,478 people) was for Higher Education and Universities. The mean number of people in each sector was 21,364. The distribution of males and females across sectors is not due to chance (Chi-square test of independence = 188380, df = 107, p < 0.001).

4.2.1. Male dominated sectors

Males dominate in 87% of the 108 economic sectors. The overall average male bias across all sectors is 62%. Males dominate by over 60% representation in approximately 65% of sectors. The top ten male-dominated economic sectors are shown in Table 9. Given that general web presence for males at the .uk level is 53%, these numbers represent serious differences in the sexes as to career choices, or career opportunities. Some well-known suspects for male dominance are evidenced: 86% male bias in Investment Banking and Advisory, Industrial Automation scores 82% males, and Oil and Energy are at 78% male. According to ONS data, about 80%

### Table 7. The .uk domain gender split from over 2.3m people detected online.

| Gender | Online Audit | ONS estimate |
|--------|--------------|--------------|
| Men    | 52.84%       | 52.97%       |
| Women  | 47.16%       | 47.03%       |

### Table 8. Gender split by .uk sub-domain.

| Domain     | Men-% | Women-% |
|------------|-------|---------|
| .police.uk | 66.20 | 33.80   |
| .ltd.uk    | 62.77 | 37.23   |
| .gov.uk    | 59.50 | 40.50   |
| .co.uk     | 57.17 | 42.83   |
| .ac.uk     | 52.38 | 47.62   |
| .org.uk    | 50.44 | 49.56   |
| .nhs.uk    | 37.19 | 62.81   |
| .sch.uk    | 25.53 | 74.47   |
of the employees in these sectors are men (ONS, 2017a,b,c), which is exactly borne out by the websites’ data.

4.2.2. Female dominated sectors

Whist over two thirds of the economic sectors are biased towards men, as is suggested by the male-female online presence of 53:47, there must be subsets of economic sectors that bias towards women. The top ten are given in Table 10. The above gives approximately 5% of economic sectors where women outnumber men, with 6 instances of women above 60% representation. Veterinary science, with over 14,504 people detected across some 615 organisations, has a 79% female dominance. This interesting figure supports data on female degree choices presented in the Global Gender Gap Report (WEF, 2017). The high level of women within academia for veterinary science suggests an early selection bias which is subsequently manifested within work roles.

Table 9. Top male dominated sectors.

| Sector                                | Women-% | Men-% |
|---------------------------------------|---------|-------|
| Investment Banking and Advisory       | 14.20   | 85.80 |
| Industrial Automation                 | 17.51   | 82.49 |
| Machinery                             | 18.05   | 81.95 |
| Mining and Metals                     | 18.89   | 81.11 |
| Mechanical and Industrial Engineering | 20.84   | 79.16 |
| Computer Networking and Security      | 21.11   | 78.89 |
| Civil Engineering                     | 21.24   | 78.76 |
| Semiconductors and Electronic Systems | 21.77   | 78.23 |
| Oil and Energy                        | 22.29   | 77.71 |
| Sports                                | 22.39   | 77.61 |

Table 10. Top female dominated sectors.

| Sector                                | Women-% | Men-% |
|---------------------------------------|---------|-------|
| Veterinary                            | 78.92   | 21.08 |
| Primary and Secondary Education       | 71.93   | 28.07 |
| Cosmetics and Toiletries              | 71.06   | 28.94 |
| Alternative Medicine                  | 67.66   | 32.34 |
| Hospitals and Medical Practices       | 66.93   | 33.07 |
| Textiles                              | 63.13   | 36.87 |
| Health, Wellness and Fitness          | 58.76   | 41.24 |
| Translation                           | 57.98   | 42.02 |
| Apparel and Fashion                   | 56.20   | 43.80 |
| Glass, Ceramics and Concrete          | 53.64   | 46.36 |
The next female-biased sectors are Primary and Secondary Education (72% women) and Cosmetics and Toiletries (71% women), followed by sectors covering activities related to Healthcare, Education or Arts. In all, women have more than 50% representation in 12.96% of economic sectors (14 of the 108).

4.2.3. Middle bias

The full data is presented in the Supplementary Table. There are many surprises. The following are representative of some middle sectors:

- Newspapers and Magazines — 64.51% males.
- Media Production — 66.22% males.
- Music — 69.20% males.
- Writing and Editing — 72.68% males.

Why the above have such high male biases seems under reported?

As can be seen in Table 11, the most evenly balanced sectors by gender are Charities and Foundations, and Leisure, Travel, and Tourism, and Universities all within reach of the general ONS gender figures.

The next line of enquiry was to examine roles: what part does gender play with respect to job roles?

4.3. Job roles

If above is presented a world of gender division, then the next line of enquiry was to reveal a chasm of representation. Just over a third of the online audit data (36.3%) could be assigned a gender and a job role. In total more than 10,000 distinct roles were identified, with a mean of 79 people per role.

The most common roles and their gender balance are shown in Table 12. Note the reversal of gender across the roles. Directors, Partners, are male, Managers are mixed.

| Sector                                      | Women-% | Men-% |
|---------------------------------------------|---------|-------|
| Charities and Foundations                   | 50.26   | 49.74 |
| Leisure, Travel and Tourism                | 49.05   | 50.95 |
| Luxury Goods and Jewellery                 | 48.91   | 51.09 |
| Government Agencies and Other Public Bodies| 48.61   | 51.39 |
| Libraries                                  | 48.59   | 51.41 |
| Higher Education and Universities          | 47.71   | 52.29 |
| Charities and Foundations                   | 50.26   | 49.74 |
and then Teachers and Assistants are female. The female-biased roles are more biased than the male roles too, so women are disproportionality clustering around these jobs — even within an already biased picture. As has been researched elsewhere, senior roles tend to be male dominated, specifically for roles such as Director and Officer (Egon Zehnder, 2016). If organisations tend to list their more senior employees online then it would be a consequence of the seniority bias that men so dominated the online space. However, once again, the fact that the total online figures revert to the ONS workplace gender split suggests that the bias is both sector career choices, as well as a seniority bias. Men and women cluster in different subsectors, and have different roles within them. Both questions were investigated further.

### 4.3.1. Leadership roles

In Table 13 are presented the leadership roles across the .co.uk domain. This shows the five of the “C-suite” leadership roles (Kreutzer et al., 2017), and Chairperson for good measure. Egon Zehnder (2016) found for the UK, 26.3% of the women hold board positions. The same study showed that women tend to occupy fewer CEO positions (7.8%) as compared to CFO positions (13.7%). In the findings women score an average of 17% for these top positions.

### 4.3.2. Support roles

Seeing the level of bias in the leadership roles, and the nature of the female under representation in large areas, a natural question to arise was where are women? Table 14 suggests they are in what are termed “support” roles (ONS, 2017a,b,c;

### Table 12. Most common online roles.

| Role        | Women-% | Men-% |
|-------------|---------|-------|
| Director    | 29.17   | 70.83 |
| Partner     | 29.49   | 70.51 |
| Manager     | 51.59   | 48.41 |
| Teacher     | 76.73   | 23.27 |
| Assistant   | 78.81   | 21.19 |

### Table 13. Leadership roles.

| Role                  | Women-% | Men-% |
|-----------------------|---------|-------|
| Chief Technical Officer | 4.53    | 95.47 |
| Chairperson           | 8.00    | 92.00 |
| Chief Finance Officer | 17.88   | 82.12 |
| Chief Executive Officer | 22.34  | 77.66 |
| Chief Operations Officer | 31.79  | 68.21 |
As can be seen in Table 14, women perform support roles at vastly disproportionate levels, on average 87% of the time compared with men. The whole area revealed by these segmentations is interesting and possibly complex. This theme was examined further by analysing gender-dominated sectors and investigating how roles are segmented within them.

### 4.3.3. Dominant sector roles: veterinary, investment banking and teaching

Veterinary science has a 78.92% bias towards women. For roles within this sector, the gender splits are shown in Table 15. The first point is that the female dominance seems to persist throughout the roles. However, in two more senior roles of Surgeon and Manager, men recover the lost ground, with a noticeable “kick” in representation.

To contrast this, for a male-dominated sector, Investment Banking and Advisory (with over 86% men), it is noticed a gradual increase in female representation as the role’s seniority decreases (Table 16). For teaching, a similar pattern of a male

| Role                | Women-% | Men-% |
|---------------------|---------|-------|
| Care Assistant      | 94.25   | 5.75  |
| Receptionist        | 94.32   | 5.68  |
| Veterinary Nurse    | 93.47   | 6.53  |
| Area/Office Manager | 71.78   | 28.22 |
| Veterinary Surgeon  | 69.56   | 30.44 |

| Role                | Women-% | Men-% |
|---------------------|---------|-------|
| CEO                 | 8.62    | 91.38 |
| Partner             | 10.81   | 89.19 |
| Director            | 12.77   | 87.23 |
| Research Analyst    | 25.00   | 75.00 |
| Associate           | 29.41   | 70.59 |

Table 14. Support roles.

| Role             | Women-% | Men-% |
|------------------|---------|-------|
| Legal Secretary  | 95.73   | 4.27  |
| Receptionist     | 94.05   | 5.95  |
| Administrator    | 83.21   | 16.79 |
| Assistant        | 78.30   | 21.70 |
| Supervisor       | 68.16   | 31.84 |

Table 15. Top roles in the Veterinary Sector.

Table 16. Top roles in the Investment Banking and Advisory Sector.
representation “kick” at governor is found and at teacher versus Assistant Teacher (Table 17).

4.3.4. Most dominated roles

This section examines the most dominated roles across all economic sectors. The male-dominated roles seem somewhat different in character than the leadership roles examined earlier (Table 18). To some degree, it could be asked: why should these roles be male? A status response, based on desirability or some level of reward, does not seem immediately obvious. It seems more likely that these jobs may be more appealing to men rather than necessarily linked to some deliberate exclusion of women, although a Service Engineer at 98% male seems extreme. Support and medical applications seem to be the most obvious characteristics across female-dominated roles (Table 19). This list seems to conform to traditional gender stereotypes.

Table 17. Top roles in Teaching-Related\(^1\) Sectors.

| Role              | Women-% | Men-% |
|-------------------|---------|-------|
| Governor          | 55.51   | 44.49 |
| Supervisor        | 74.72   | 25.28 |
| Teacher           | 78.92   | 21.08 |
| Teaching Assistant| 87.88   | 12.12 |

\(^1\)Includes Higher Education and Universities and Primary and Secondary Education.

Table 18. Most male-dominated roles.

| Role              | Women-% | Men-% |
|-------------------|---------|-------|
| Service Engineer  | 2.01    | 97.99 |
| Technical Director| 4.33    | 95.67 |
| CTO               | 4.53    | 95.47 |
| Warehouse Manager | 5.06    | 94.94 |
| Engineering Manager | 5.47  | 94.53 |

Table 19. Most female-dominated roles.

| Role              | Women-% | Men-% |
|-------------------|---------|-------|
| Medical Secretary | 96.68   | 3.32  |
| Practice Nurse    | 96.20   | 3.80  |
| Dental Nurse      | 95.93   | 4.07  |
| Legal Secretary   | 95.73   | 4.27  |
| Dental Hygienist  | 95.03   | 4.97  |

https://doi.org/10.1016/j.heliyon.2018.e01001

2405-8440© 2018 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
4.4. The NHS and the BBC

To conclude the research two major UK institutions and their online representations are highlighted. The gender breakdowns for the NHS and the BBC are shown in Table 20.

4.4.1. The BBC

The BBC is in the bottom quarter for female representation as shown in this comparison (Table 21).

The BBC concedes 7% in female representation against its peer group. With a sample of 5762 people found on the BBC website this suggests bias even over a biased group. By contrast, the BBC reports 47.7% women in employment and 43.3% in leadership (BBC, 2018).

The leadership theme seems to be well presented across most prominent BBC job titles, with strong male bias for Directors, Presenters and even Performers (Table 22). The BBC provides some status, and having a “profile” online could be seen

Table 20. Gender representation online.

| Institution | Men-% | Women-% |
|-------------|-------|---------|
| NHS         | 37.19 | 62.81   |
| BBC         | 72.53 | 27.47   |

Table 21. BBC comparison with related sectors.

| Sector                       | Women-% | Men-% | Rank |
|------------------------------|---------|-------|------|
| BBC                          | 27.47   | 72.53 | 84   |
| Media Production             | 33.78   | 66.22 | 66   |
| Newspapers and Magazines     | 35.49   | 64.51 | 57   |
| Broadcast Media (TV, Radio)  | 35.89   | 64.11 | 55   |

Table 22. Most frequent roles on the BBC websites.

| Role     | Women-% | Men-% |
|----------|---------|-------|
| Composer | 12.32   | 87.68 |
| Director | 21.88   | 78.13 |
| Presenter| 22.64   | 77.36 |
| Performer| 27.47   | 72.53 |
| Producer | 44.44   | 55.56 |

1 Director role refers to a director of a play.
2 Producer role was determined as an aggregation of producer-related job roles (e.g. series producer, planning producer).
as an extension of such status, and may mean these roles are more keenly contested. The fact that 12% of Composers (as in musical composers) are female, from a population of 138, seems a legacy of long-standing gender bias, but it may be something that the institution should be aware of.

4.4.2. The NHS

The NHS sits as follows against its peer groups (Table 23).

The NHS on this reckoning may be slightly less biased than its peer group, although well below government agencies. NHS reports that 77% of its workforce are women. However, just 46% senior manager roles in the NHS are held by females (NHS, 2018).

The medical profession captures a considerable bias towards women, as seen in the most dominated roles section presented earlier. Table 24 shows how this maps on to roles for the NHS. Nursing is one of the most biased roles. However, the NHS is doing better (even at 87%) than over the medical sectors as a whole. Still the extreme position stands out.

The most surprising figure above seem to be the Surgeon role, with over 76% males. This must indicate a subculture within the medical profession, which the professional bodies may wish to address. Also seen in the roles listing is the recovery to men in leadership positions. In other words, men seem to recover the lost ground when looking at more senior roles.

Table 23. NHS comparison with related sectors.

| Sector                        | Women-% | Men-% | Rank |
|-------------------------------|---------|-------|------|
| NHS                           | 62.81   | 37.19 | 6    |
| Hospitals and Medical Practices| 66.93   | 33.07 | 5    |
| Health, Wellness Fitness      | 58.76   | 41.24 | 7    |
| Government Agencies and Other Public Bodies | 48.61   | 51.39 | 17   |

Table 24. NHS Online positions.

| Role            | Women-% | Men-% |
|-----------------|---------|-------|
| Surgeon/Consultant | 23.92  | 76.08 |
| Director        | 48.02   | 51.98 |
| Manager         | 70.84   | 29.16 |
| Secretary       | 75.63   | 24.37 |
| Nurse           | 87.18   | 12.82 |
5. Discussion

At the top level, the number of men and women found online conforms to the workplace figures, but what is extraordinary is the level of separation below this. Sectors can be biased one way or another, but the vast majority, are biased towards men. There seems to be a general career selection process taking place to segment gender. In line with other academic findings, women are over-represented in sectors related to health, social work and education (McGuinness, 2018). In a small group of 5% of sectors, there is female domination (over 60%) and an equally small group there is a broadly gender-neutral representation.

Another phenomenon discovered online is the bias within senior and support roles. The data strongly support the view presented by the ONS that men dominate leadership positions, and that women embody support and facilitation roles (ONS, 2013). Even in the case of female-dominated sectors, a pattern of rising male representation in leadership positions is seen.

Thus, gender representation online shows widespread biases. It reveals a form of gender separation in which men and women seek different career paths. For leadership, the numbers speak for themselves. If being a CEO is a coveted position, imagine a life test that says: there 82% chance of failure, just based on gender? The high level of women in support roles also seems complex. Are males refusing these jobs? Do females accept them, or out-compete for them? The latter seems unlikely. This level of segregation suggests some form of cultural contour that possibly starts very early in life, it may even be cumulative clustering of peer groups, early school choices, social groups and cultural re-enforcement, as well as prejudice.

The separations of roles suggest some jobs are seen as female and others as male, with genders possibly self-selecting to reinforce these groupings. One side effect of the prominence is an online normalisation of representation in certain roles. In one sense people are being advertised in their roles by the way organisations choose to show them on their websites.

What is also possible within support roles is that they offer non-leadership leading career paths. Can you move from Receptionist to CEO? If these roles do progress to leadership, then somehow men seem to be avoiding them, as men seem to skip these pathways. The most obvious conclusion seems that they are bounded roles, with gender bias, that then exclude top positions, and men avoid them, and/or are excluded from them.

All of the above discussion for areas of analysis, and possible explanation, are way beyond the scope of this research and are the subject of deep and widespread ongoing research (WEF, 2017; British Council, 2016). This study simply represent more data to support this ongoing work.
Finally, it is also recognized that this study has a number of limitations. Firstly, the study does not distinguish whether the people are actually employed by the organisations in question. Second, the web data can be characterised by a seniority bias as specific organisations and industries (e.g. Construction) tend to represent only their more senior staff. Lastly, web data is dynamic and the results could also be extended to measure variance in the representations found and should therefore be recognised as a sample of data from 2017.

5.1. Comment

One aspect that is also worth comment, is that the data gathered relates to individuals. That is each person counted was harvested from open web research. No permission was sought or obtained. What are the ethics of this? A particular attention has been given to the data used, from organisations websites, with little to no "political" stance implied. But the prospect of individuals being tracked, or researched, from the open internet is here. How such data should be treated in the context of “desktop” global surveillance is clearly now a tangible issue. To coin a phrase, “little-brother, or little sister… little-person?” is watching you. How this technology should be used and can be used is open, and with recent controversy with respect to social media data and political advertising, this issue introduces itself within the realm of research.

6. Conclusion

To our knowledge, this is the first systematic “big data” analysis of social and economic data of this scale against the open web. The results match the official statistics and reflect most of the existing gender differences across industries and job roles. This paper has shown that web data is rich in insights and possibilities, clearly offering a scale of analysis previously impossible, and solving the main challenges brought by surveys: time and cost. Much more work should be done on this data. First, a deeper comparison between web and real-world data should be carried out (in terms of sector membership and gender representation). Second, future work should involve analysing the dynamics of people and gender on the web across industries, and examine to what extent the changes influenced by the real-world events.

Declarations

Author contribution statement

Ana-Maria Huluba: Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.
Jason Kingdon: Conceived and designed the experiments; Analyzed and interpreted the data; Wrote the paper.

Iain McLaren: Conceived and designed the experiments; Contributed reagents, materials, analysis tools or data.

**Funding statement**

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**Competing interest statement**

The authors declare no conflict of interest.

**Additional information**

Supplementary content related to this article has been published online at [https://doi.org/10.1016/j.heliyon.2018.e01001](https://doi.org/10.1016/j.heliyon.2018.e01001).

**References**

Askitas, N., Zimmermann, K., 2015. The Internet as a Data Source for Advancement in Social Sciences. IZA Discussion Papers(8899).

Baumgartner, R., Frölich, O., Gottlob, G., Harz, P., Herzog, M., Lehmann, P., Wien, T., 2005. Web data extraction for business intelligence: the lixto approach. In: Proc. 12th Conference on Datenbanksysteme in Büro, Technik und Wissenschaft, pp. 30–47.

BBC, 2018. Bbc Equality Information Report 2017/2018. Retrieved from. [http://downloads.bbc.co.uk/diversity/pdf/bbc-equality-information-report-2017-18.pdf](http://downloads.bbc.co.uk/diversity/pdf/bbc-equality-information-report-2017-18.pdf).

Benfield, J.A., Szlemko, W.J., 2006. Internet-based data collection: promises and realities. J. Res. Pract. 2 (2). Retrieved from. [http://jrp.icaap.org/index.php/jrp/article/view/30/51](http://jrp.icaap.org/index.php/jrp/article/view/30/51).

BLS, 2018. Occupational Outlook Handbook. Retrieved from United states Department of Labour: [https://www.bls.gov/ooh/office-and-administrative-support/home.htm](https://www.bls.gov/ooh/office-and-administrative-support/home.htm).

Bollacker, K., Evans, C., Paritosh, P., Sturges, T., Taylor, J., 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, pp. 1247–1250.
British Council, 2016. Gender equality and Empowerment of Women and Girls in the UK. Retrieved from: https://www.britishcouncil.org/sites/default/files/gender_equality_an_empowerment_in_the_uk_0.pdf.

Daas, P.J., Puts, M.J., Buelens, B., van den Hurk, P.A., 2015. Big data as a source for official statistics. J. Off. Stat. 31 (2), 249–262.

Dong, X., Gabrilovich, E., Heitz, G., Horn, W., LN, Murphy, K., Zhang, W., 2014. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 601–610.

Egon Zehnder, 2016. Global Board Diversity Analysis. Egon Zehnder. Retrieved from: https://www.egonzehnder.com/GBDA.

Fast, E., Vachovsky, T., Bernstein, M.S., 2016. Shirtless and Dangerous: Quantifying Linguistic Signals of Gender Bias in an Online Fiction Writing Community. ICWSM, pp. 112–120.

Ferrara, E., De Meo, P., Fiumara, G., Baumgartner, R., 2014. Web data extraction, applications and techniques: a survey. Knowl. Base Syst. 301–323.

Francoeur, C., Labelle, R., Sinclair-Desgagné, B., 2006. Gender diversity in corporate governance and top management. J. Bus. Ethics 34.

Goldin, C., 2014. A Pollution Theory of Discrimination Male and Female Differences in Occupations and Earnings. Retrieved from: http://www.nber.org/chapters/c12904.pdf.

GOV.UK, 2018. Report Your Gender Pay gap Data. Retrieved from: https://www.gov.uk/report-gender-pay-gap-data.

Haranko, K., Zagheni, E., Garimella, K., Weber, I., 2018. Professional Gender Gaps across US Cities. arXiv preprint.

Hoogendoorn, S., Oosterbeek, H., van Praag, H., 2013. The impact of gender diversity on the performance of business teams: evidence from a field experiment. Manag. Sci.

Huws, U., 2014. Labor in the Global Digital Economy: the Cybertariat Comes of Age. Monthly Review Press, New York.

Jia, S., Lansdall-Welfare, T., Cristianini, N., 2015. May. Measuring gender bias in news images. In: Proceedings of the 24th International Conference on World Wide Web. ACM, pp. 893–898.

Jones, J., 2013. UK Service Industries: Definition, Classification and Evolution. Office for National Statistics, London.
Kay, M., Matuszek, C., Munson, S.A., 2015. Unequal representation and gender stereotypes in image search results for occupations. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 3819–3828.

Kreutzer, R.T., Neugebauer, T., Pattloch, A., 2017. Digital Business Leadership. Springer Gabler.

McGuinness, F., 2018. Women and the Economy. House of Commons Library.

Mockapetris, P.V., 1983. Domain Names: Implementation Specification. https://www.rfc-editor.org/rfc/rfc/rfc883.txt.pdf.

Nathan, M., Rosso, A., 2015. Mapping digital businesses with big data: some early findings from the UK. Res. Pol. 1714–1733.

NHS, 2018. Gender in the NHS. Retrieved from. https://www.nhsemployers.org/-/media/Employers/Images/2018-D-and-I-infographics/Gender-in-the-NHS-2018.pdf.

Niu, F.Z., 2012. Elementary: large-scale knowledge-base construction via machine learning and statistical inference. Int. J. Semantic Web Inf. Syst. 8 (3), 42–73.

ONS, 2013. Women in the Labour Market. Retrieved from. https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandemployeetypes/articles/womeninthelabourmarket/2013-09-25.

ONS, 2017a. Annual Survey of Hours and Earnings: 2017 Provisional and 2016 Revised Results. Retrieved from. https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/bulletins/annualsurveyofhoursandearnings/2017provisionaland2016revisedresults.

ONS, 2017b. E-commerce and ICT Activity: 2016. https://www.ons.gov.uk/businessindustryandtrade/itandinternetindustry/bulletins/eCommerceandictactivity/2016.

ONS, 2017c. EMP04: Employment by Occupation. Office for National Statistics.

Otterbacher, J., Bates, J., Clough, P., 2017. Competent men and warm women: gender stereotypes and backlash in image search results. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 6620–6631.

Paulheim, H., 2017. Knowledge graph refinement: a survey of approaches and evaluation methods. Semantic Web 8 (3), 489–508.

Smith, G.J., 2007. Internet Law and Regulation. Sweet & Maxwell.

Strazny, P., 2005. Encyclopedia of Linguistics. Taylor and Francis Books, Inc.
Suchanek, F.M., Kasneci, G., Weikum, G., 2007. Yago: a core of semantic knowledge. In: Proceedings of the 16th International Conference on World Wide Web, pp. 697–706.

WEF, 2017. The Global Gender Gap Report. World Economic Forum. Retrieved from. https://www.weforum.org/reports/the-global-gender-gap-report-2017.

Yan, S., Ge, C., 2016. Gender Difference in Competition Preference and Work Duration in the IT Industry: Linkedin Evidence.

Zagheni, E., Weber, I., 2015. Demographic Research with non-representative internet data. J. Manpow. 36 (1).

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S., 2016. Quality assessment for linked Data: A survey. Semantic Web 7 (1), 63–93.