Traces of Unequal Entry Requirement for Illustrious People on Wikipedia Based on their Gender

Lea Krivaa

CS Department, IT University of Copenhagen, Denmark

Michele Coscia

CS Department, IT University of Copenhagen, Denmark

mcos@itu.dk

Wikipedia is a widely used tool people use to gather knowledge about the world, causing it to have a vast impact on the way individuals perceive the reality they live in. It is then of paramount importance that the picture of the world Wikipedia provides is accurate. We cannot afford such an important tool to eschew inclusiveness or a fair representation of reality: an inaccurate picture of the world in such a tool can be used to claim unjust and unfair positions – such as that women are inferior to men – as if they were facts, because they are enshrined on an encyclopedia. In this paper, we study issues of fair gender representations for people in history noted by multiple language editions of Wikipedia: are women underrepresented on Wikipedia? We do so via a combination of natural language processing and network science. Our results indicate that there is indeed a higher bar for women to have their own biographical page on Wikipedia: women are only included when they have more significant connections than men to the rest of the network. There are visible effects of the initiatives Wikipedia is taking to fix this issue, showing that the gap is narrowing, which validates our interpretation of the data.

1. Introduction

Wikipedia is a widely used tool, being the 7th most visited website on the Internet [3]. Wikipedia provides a widely read crowd-sourced encyclopedia that is used to gather information about the reality we live in. As a result, it has become one of the most influential websites in shaping our collective perspective of the world.

Given its impact, it is important that Wikipedia does not offer a distorted picture of reality. One axis of investigation is asking: are the facts stated in its articles factually correct? Researchers have spent a considerable effort investigating hoaxes [20], conspiracy theories [32], vandalism [5, 40], trolling [38], accuracy [9], and disinformation [34] on Wikipedia, as well as an overall comparison with other models of managing an encyclopedia [14]. A lot of work has been dedicated on methods to detect and correct misinformation [28] and vandalism [29], creating a reliability database [45]. Researchers have found that controversial topics might trigger "edit wars", whose resolution might affect the reliability of the articles [46].

However, being factually correct is not the only important issue. One could build an encyclopedia in which all the information present is accurate, but sys-

tematically omitting the contributions from a certain class of people and/or information about specific topics. Therefore, researchers have also heavily investigated Wikipedia's issues of inclusiveness and representation [25]. The research focuses on several axes, spanning from representation in the readers [18], the editorial team [21] and processes through which contributors update Wikipedia [10, 6]; as well as how Wikipedia's content covers different cultures [12], ages [35], geographical areas [17], gender identities [31, 16], and languages [26, 33]. This involves not only whether an article is included or not, but also how it is discussed in the text [42, 39].

In this paper, we focus on the issue of gender representation when it comes to historical figures included in Wikipedia. The general research question we are interested in answering is: are women underrepresented on Wikipedia? This question is broad and can be interpreted and approached in multiple ways, so we restrict our interest to the following sub-questions, specifically using a network science methodology:

(1) Are there structural differences in how the Wikipedia pages of notable men and women are connected by hyperlinks?
(2) If there are, can we use such structural differences to infer that there is a higher bar of entrance for women to have a Wikipedia biography than it is for men?
(3) Is Wikipedia acting on this discrimination and is the action effective?

Past studies show that there is a higher requirement for women to be noted than for men. In one study, authors find that the coverage of women is more fair for ancient and contemporary times, but it reaches a low point in the XIX century [19].

The shortcoming of this study is that it does not provide a target indicating what a fair gender inclusion would look like on Wikipedia. This issue is solved in other papers which find a variety of effects. Among pages of sociologists [4] there is a significant amount of exclusion that is not found, e.g., among musicians [43]. Among politicians, while inclusion between genders is comparable, articles about women politicians are significantly different, including more details about the politician's private life [30]. The shortcoming in this collection of works is the narrow scope: the models require rich data and so they focus on a smaller subset of people. As a consequence, we lack the general picture we seek in this work.

With our network science approach, we can find a way to discuss quantitatively about the inclusion criteria with a broader scope by looking at the overall structure of what is present in Wikipedia.

Here, we focus on the 1750-1950 period to consider a period with consistent record keeping and availability. We model Wikipedia's biographies as a complex network, connecting the people who have a Wikipedia page in multiple languages if their biographical pages link to one another. We also use natural language processing techniques to add edge weights to these connections. The use of network analysis is crucial because it provides us with a way to test for the statistical significance

of the edge weights. This is a key factor to address our research question, because finding a gap in the statistical significance of the edges for two groups allows us to infer disparities in the inclusion criteria for the two groups – disparities that are not simply about having more or fewer hyperlinks, but are related to the network topology itself – as we show in a simple model.

Our findings lead us to the following answers to our research questions:

(1) There are structural differences between men and women in the network structure, the most significant of them being that women's biographies tend to have more significant edge connections.
(2) We can show how these systematic differences can be linked to the fact that a woman's profile is only added if it clears a more demanding significance threshold in their connections with other profiles on Wikipedia.
(3) Wikipedia has put in place initiatives to counteract this gender representation bias. Our data shows that they are going in the right direction.

As an example for the last point, we investigate the popular initiative to add more women pages in the month of March. We indeed see a smaller gap for pages that were added in that month. Our observation is not only in accordance with the literature [22, 41], but it also represents a validation of our approach: efforts to reduce the gender representation gap leave a noticeable trace in our data, in accordance to our interpretation of it. Incidentally, the gap – while smaller – remains also in the month of March, which is also consistent with the literature (one proposed explanation is that the biographies added during these initiatives are more likely to be subsequently marked for deletion [41]).

The data and the code we use to generate our findings is publicly available to reproduce them (`https://www.michelecoscia.com/?page_id=2285`).

2. Data

2.1. *Data Collection*

We use as seed list the list of people from Pantheon [47, 7] (retrieved on November 24th, 2021), which includes a total of 88,937 individuals across all human history. We use Pantheon data because it comes with a pre-curated list. Specifically, Pantheon data only includes people who have a Wikipedia page in multiple languages, which implies they have attracted the collective attention across a wider population than the one speaking a single given language.

Pantheon also already provides the information about the gender of the person, which we do not need to retrieve ourselves. We find that 84,463 out of 88,937 have a gender information that maps to either male or female and we focus on these nodes for our analysis.

We collect the content of the Wikipedia page with a crawling process on February 12th, 2023. For each page, we collect the main text and all the hyperlinks the page contains. In some cases, the Wikipedia page id did not lead to a crawlable page –

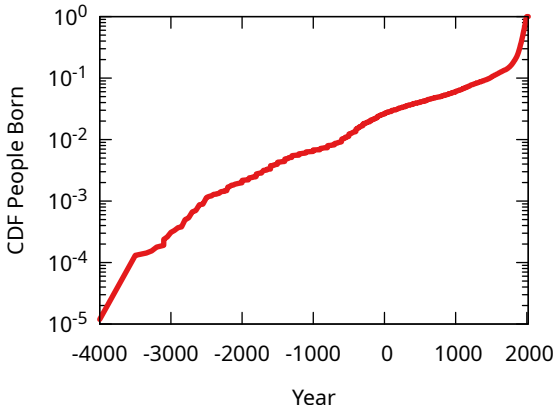

Fig. 1. The cumulative count of people (y axis) born in a given year (x axis) in the Pantheon dataset.

possibly the slug was modified as the result of a new entry. This results in a minor loss of nodes in our network.

2.2. *Data Cleaning*

Other works on notable people using a wider scope than ours [36, 23] have noted a discontinuity around the XVIII century in the patterns of deaths of notable people. Moreover, we observe that in our selected Pantheon dataset the record keeping for contemporary people seem to be more lax than for the past.

Figure 1 somewhat confirms this impression. The first 5750 years in the data (until 1750) contain less than 15% of the births. On the other hand, there is a noticeable post-1950 spike – the last 70 years include nearly 46% of all people in the Pantheon data.

For this reason, we stick with a somewhat arbitrary choice of including only individuals in the 1750-1950 period, to try focusing on a relatively long period with relatively consistent admission criteria. This results in a set of 32,901 people.

We then create the basic graph by having each person as a node. We connect two nodes if there is at least one hyperlink between the two pages. Note that we keep edge directions, so person u connecting to person v does not necessarily mean that there is a connection also going from v to u. We remove self-loops and multiple parallel edges. If a node does not have any hyperlink neither incoming nor outgoing – or we cannot retrieve its biography from Wikipedia due to URLs moving over time –, we drop it from the network.

This results in a network with 9,540 nodes and 191,803 edges. For each node we have the full text of the Wikipedia biography, which is what we use to estimate its edges' weights via an NLP algorithm that we describe in the Methods section.

Note that there is a considerable amount of edges (149,681, 78%) with weight

Statistic	With 0-weighted edges		Without 0-weighted edges	
	M	F	M	F
Node Count	8,177	1,363	7,341	1,235
AVG Outdegree	20.3	18.9	4.84	5.32
AVG Indegree	20.6	17.0	5.05	4.08

Table 1. Some summary statistics divided by gender and by network type.

equal to zero. This is due to the fact that two pages might be linked by a hyperlink that does not appear in the main text of a biography. We decide to count this as zero weight rather than having a weight of one, because such links are usually devoid of semantic meaning: they often appear in information boxes and do not genuinely express a relationship. For instance, Aristotle (`https://en.wikipedia.org/wiki/Aristotle`) is connected to Hu Shih (`https://en.wikipedia.org/wiki/Hu_Shih`) merely on the basis of both being literary theorists. We think considering this as a meaningful relationship is dubious.

2.3. *Exploratory Data Analysis*

Here we consider the networks with and without zero weighted edges as different: we provide summary statistics about both, but we consider the network without zero weighted edges as our primary target for the analysis.

Figure 2 shows the adjacency matrix of the network we analyze, including only non zero weighted edges. The network has primarily a nested structure, a sign of a core-periphery organization – which one can appreciate noticing how much more dense the top left corner of the matrix is, and how empty the bottom right corner is. We can see there is a trace on the diagonal, hinting at the possible existence of some small secondary clusters that could be identified in future works.

Table 1 summarizes node counts and average degrees for all node types in all networks we consider. Most nodes in either network are male. While the disparity is large, this is not by itself evidence of a fault in Wikipedia: past record suppression – regardless whether the suppression was caused by not recording achievement by women or by historians ignoring existing records – can be the cause of this disparity. If the records are not accessible, then Wikipedia's editors cannot fix this specific representation gap.

In both networks, out-degree values between genders are closer than in-degree values. In further sections, we show that the differences in degrees can be explained by homophily – the tendency of connecting with entities similar to oneself –: if there are more male nodes to connect to, male nodes should obtain a higher degree since they preferably connect to them. However, homophily should affect in- and out-degree equally, but we see that this is not the case, showing a potential link-making disparity in Wikipedia: when writing about men it is less likely to point to a woman than vice versa.

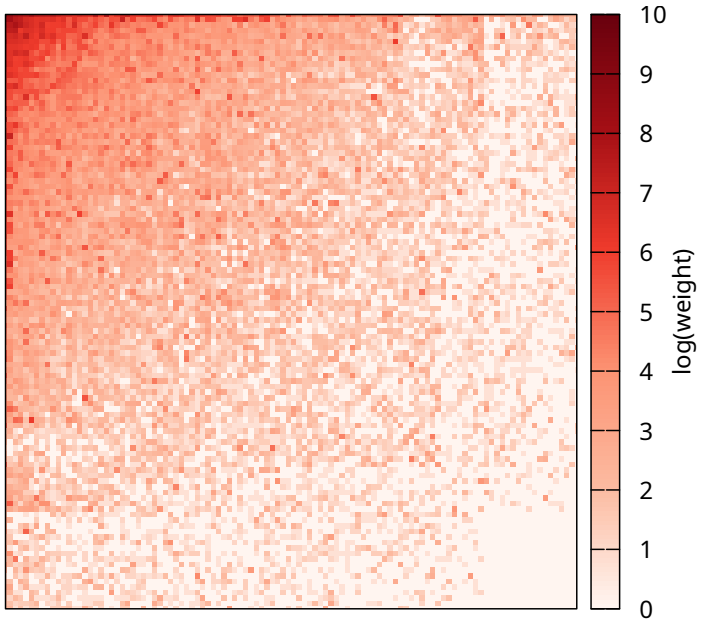

Fig. 2. The adjacency matrix of the network, with each square representing the total edge weight between the nodes in a given cell from high (dark) to low (bright). The rows and columns of the matrix are sorted by total weight of a node's connection, in descending order.

The degree distributions (Figure 3) show that, for men, there is an expected difference between in- and out-degree – out-degree is limited by article length, while an individual can be pointed by an arbitrary number of other pages. Surprisingly, the in- and out-degrees for women does not show this expected gap. By manually looking at page creation dates for a handful of randomly selected women, we spotted that, in general, women tend to point more to already-existing man pages and, additionally, in this case it is less likely for a woman to get a link back. We plan to investigate this further in future works.

The lone noticeable exception when it comes to women's in-degrees is Queen Victoria, which is an outlier for a given in-degree value given the number of women in the dataset. However, the highest in-degree node is still a man (Napoleon with zero weights, Hitler without).

Interestingly, when ignoring zero weighted edges there is no difference in the

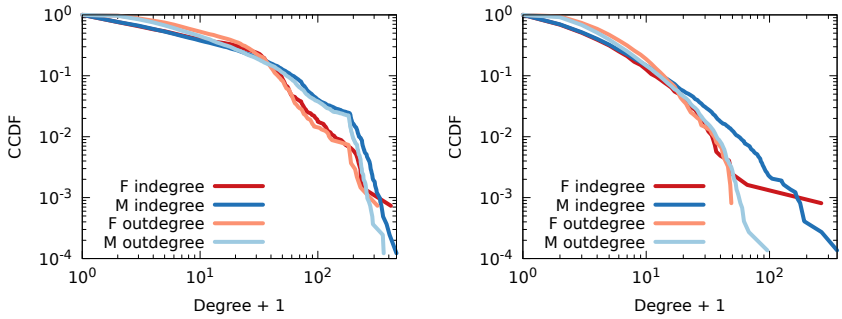

Fig. 3. The probability (y axis) a node has a given degree (x axis) or higher. Colors encode the node type, based on gender. (Left) Network with zero weighted edges, (Right) network without zero weighted edges.

out-degree distributions between males and females. The highest out-degree nodes are also men in both networks (Ralph Waldo Emerson with zero weights, Ulysses S. Grant without), while the highest out-degree women are Germaine de Staël with zero weights, and Princess Royal Victoria without.

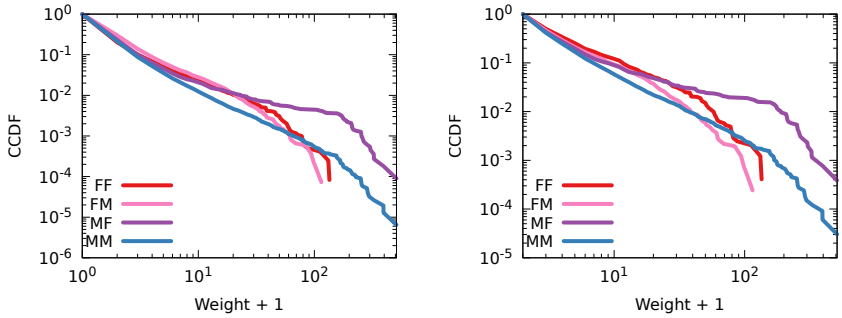

Fig. 4. The probability (y axis) an edge has a given weight (x axis) or higher. Colors encode the edge type, based on gender. (Left) Network with zero weighted edges, (Right) network without zero weighted edges.

Finally, edge weight distributions (Figure 4) show that edges originating from men tend to have higher weights, although edges between two men are less likely to have mid-weighted edges than edges involving at least one woman. The only difference with excluding zero weighted edges is in censoring the distribution for that specific value.

3. Results

The main objective of this paper is to find disparities in how women are treated differently from men on Wikipedia when it comes to linking them in a network

Edge Type	With 0-weighted edges	Without 0-weighted edges
MM	0.622	0.269
MF	0.576	0.340
FM	0.468	0.214
FF	0.716	0.258

Table 2. The reciprocity values for each edge type for networks both including and excluding zero weighted edges.

describing human history. The previous section highlighted some differences, but those can have trivial explanations – e.g. they could be due to the expected difference in numbers of male versus female nodes and the consequent homophily.

We organize our results as follows: first we want to verify the homophily hypothesis, then we propose a more principled analysis of the disparities in connections by taking into account the statistical significance of the edges weights. Finally, we conclude with an attempt to connect the differences we found with initiatives in the Wikipedia editorial team.

3.1. *Network Statistics*

We now look at some gender discrepancies according to several network measures. We perform our analysis on the complete network. We include in Supplementary Material Section 1 a temporal analysis showing the evolution of these measures over time.

In the network there is no significant difference in the centrality of men and women. The average PageRank is comparable and the disparity in top ranking nodes can be explained with the imbalance in the node counts between the two genders.

There is a gender difference in the average clustering coefficient, but it is small – the men and women average clustering coefficients are 0.392 and 0.402 in the network with zero-weighted edges, and 0.212 and 0.246 in the network without zero-weighted edges, respectively.

We decide to focus the assortativity angle. The assortativity coefficient is 0.419 for the whole network and 0.331 if we remove the zero weighted edges. This means that men tend to connect to men and women to women. Given the lower node count for women, this could explain the disparity.

This is not conclusive, because the assortative coefficient is a network property, but male and female node might still behave differently – and the edge's direction might play a role. To investigate this, we make an analysis of edges distinguishing them into types (MM, MF, FM, FF).

First we look at reciprocity, which is the share of links that point in both directions. Table 2 shows that, in the network with zero-weighted edges, there is a homophilic tendency to reciprocate – same gender links are more likely to point

Edge Type	With 0-weighted edges		Without 0-weighted edges	
	Null Expectation	Observation	Null Expectation	Observation
MM	0.7346	0.8077	0.7327	0.7826
MF	0.1225	0.0580	0.1233	0.0615
FM	0.1225	0.0714	0.1233	0.0978
FF	0.0204	0.0629	0.0207	0.0581

Table 3. The expected and observed probabilities for each edge type for network both including and excluding zero weighted edges.

back, and more so for women. Cross-gender links show that it is much more likely to reciprocate links pointing from men to women than vice versa, showing a gender imbalance that can explain the observed differences in the degree distributions. While the homophilic reciprocity goes away when we ignore zero-weighted errors, the gender imbalance in reciprocity stays.

We expand this analysis by comparing the observed edge counts of all types with what we would expect them to be if nodes connected randomly – given the count disparity between M and F nodes. We can estimate what should be the probability of observing an edge of a given type (MM, MF, FM, or FF) via a hypergeometric null model. The expected edge frequencies are equivalent to the probability of extracting zero, one, or two F nodes with two attempts (an edge) following a hypergeometric distribution. We use the hypergeometric distribution, because this is extraction without replacement, given that we disallow self loops.

Table 3 compares these probabilities with the observed edge counts, for both the network with and without zero weighted edges. The table confirms the assortativity coefficients, with MM and FF nodes overrepresented – more so in the network with zero weighted edges.

However, when taking the direction into account, we see a significant disparity between genders. The null model expects MF and FM edges to be equally likely. However, that is not what we observe in the network, where F nodes are more likely to connect to M nodes than vice versa. This is true regardless whether we include or exclude zero weighted edges.

Here we found a disparity that cannot be explained simply by the higher male node count. There is indeed a difference between men and women when it comes to create their edges on Wikipedia, and in the rest of the paper we set to further explore it by taking into account edge weights.

3.2. *Significance Disparities*

Applying a backboning threshold means to throw away from the network all edges that do not meet a given significance threshold p – as we detail in the Methods section. Here, p means the probability that the edge weight could be zero, just like a regular p-value in hypothesis testing.

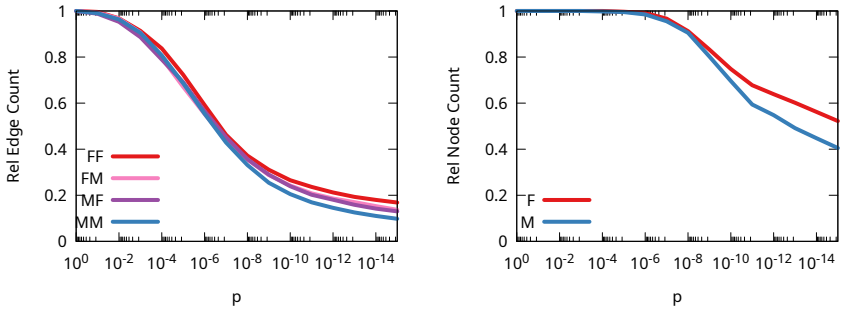

Fig. 5. The relative counts of nodes and edges (y axis) as we increase the statistical significance p (x axis) by category dependent on gender (color). (Left) Counts of edges. (Right) Counts of nodes with at least one connection in the network.

We need to test for significance for two reasons. First, our edge weight inference process is not perfect, there could be errors, and we need to take this into account. Second, even if the edge weights were perfectly estimated, we should still test their statistical significance for a given node when removing edges. Otherwise, we would only remove the edges with a smaller weight, making the filter not a proxy of the importance of the connection, but simply of the length of a biography – since longer biographies have more chances to generate a higher edge weight by mentioning a target node more times.

We would not expect differences in edge count drops between genders as we demand for a higher statistical significance, because there is no reason for people of different genders to have significantly different edge weights.

Figure 5 shows that this expectation is actually not reflected in the data we have. First, on the left, we see that FF edges tend to have higher significance. MM edges have the lowest and mixed gender edges are relatively equal – in between the single-gender ones.

The difference in edge significance is small but, when combined with the observed homophily, can lead to a large effect in terms of nodes. On the right of Figure 5 we see how the nodes evolve as we filter the network more and more. Specifically, we keep track of the nodes that have at least one surviving edge – and are thus not eliminated from the network. At very high levels of significance, the difference is stark: for $p = 10^{-15}$ we preserve more than 50% of the F nodes (52% in fact) but only 40% of the M nodes – we choose to stop at $p = 10^{-15}$ because lowering further would result in random noise added by low machine precision in representing such low numbers.

This result suggests that there is a higher bar for women to be included in Wikipedia – a suggestion that is plausible given a synthetic experiment we include in Supplementary Material Section 2. When women are included, they show their greater importance in the network by surviving the significance filtering. We should

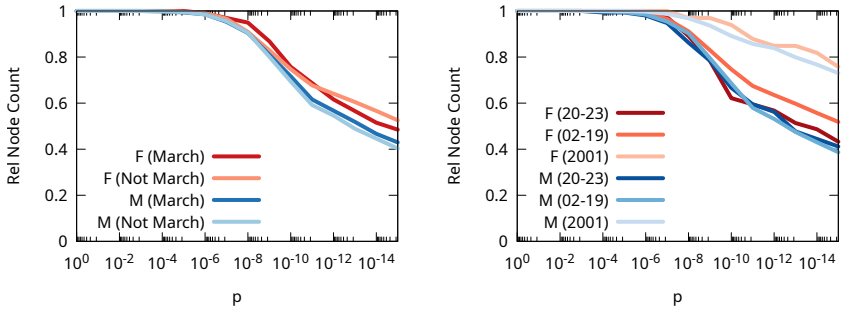

Fig. 6. The relative counts of nodes (y axis) as we increase the statistical significance p (x axis) by category dependent on gender (color). (Left) Comparing March (dark hue) with the other eleven months. (Right) Comparing different periods (hue darkness).

expect no difference in the significance pruning between M and F nodes.

3.3. *Disparities in Context*

Our explanation of the disparity is that women face a higher bar of entrance than men on Wikipedia. We can test this explanation if we find a time period in which we should expect the bar for women should be lower than normal. Luckily, there is such a period: Wikipedia "started an annual tradition in March of creating and improving Wikipedia articles about women" [1]. Therefore, we should expect the bar for entry on Wikipedia for women to be lower in March, at the very least the disparity with men should be smaller.

For each page in Wikipedia we know the exact date in which it was added. Therefore, we can keep track of the relative number of nodes added in a specific month that retain at least one link in our network at a given edge significance threshold filtering. In practice, we replicate Figure 5 (Right), but only for the month of March.

Figure 6(Left) shows the result. We see that our explanation is supported by the empirical evidence: the month of March shows that the treatment of women and men on Wikipedia is fairer – the gap is smaller than overall. While at $p = 10^{-15}$ the gap in the other eleven months is 12.3%, in March it is only 5.5% – less than half.

Another partial confirmation of our explanation could come by looking at the evolution of Wikipedia over the years – under the assumption that in recent years there have been stronger effort to reduce the disparity. Figure 6(Right) shows some promising patterns.

In 2001, the bar of inclusion was very high for both men and women, which makes sense as that was the year in which Wikipedia was created and therefore the editors needed to prioritize only the most notable people. In the 2001-2019 period, the entry bar was low for men and high for women, as issues of gender representation

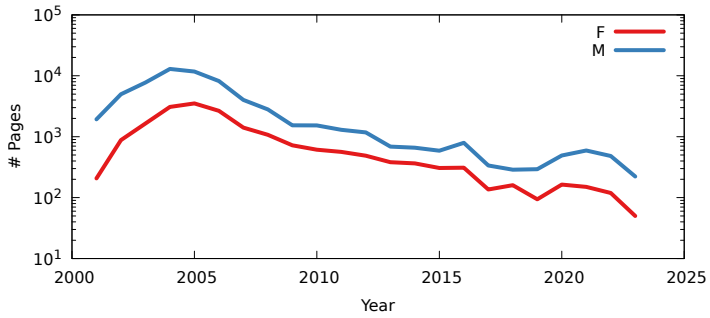

Fig. 7. The number of biographies (y axis) added to Wikipedia included in the Pantheon dataset per year (x axis) by gender (line color).

were not on the spotlight in that period. If we then look at the post 2019 period, the gap has closed: now the inclusion bar for women and for men appears to be the same.

In fact, from Figure 7 we can see that in 2001 there were only a bit more than 2,000 biographies added overall on Wikipedia – compared to 16,000 in 2004 – confirming that in 2001 it was difficult for a person to get their own Wikipedia entry. The gender ratio was also heavily skewed: less than 10% of 2001's biographies were of women, compared to 25% in 2020.

However, we should not over-interpret this result because there is a potentially alternative explanation: reversion to the mean. Wikipedia editors have simply ran out of men to add to the website – after all, the 1750-1950 period cannot really produce new individuals – and only women are left over. While we include this last result because we find it suggestive, more research is needed to make sure it really supports our point.

4. Discussion

In this paper, we investigate the disparity in treatment between men and women who have a Wikipedia page in multiple languages. We focus on admission criteria and we find that there is evidence for a disparity in treatment. When we model Wikipedia as a weighted network, connecting people via hyperlinks whose weight is the number of references in the article, we find that women tend to have more statistically significant links.

We interpret this fact as representing a higher entrance bar for women in Wikipedia: a woman is added on Wikipedia only when she has stronger connections to the existing structure than a man. We find support to this interpretation by looking at biographies added to Wikipedia in March – a month in which the editors make a conscious effort to improve female representation –: the gap in edge significance between men and women added to Wikipedia in March is smaller. Temporal patterns also partially support this interpretation: we see that biographies added in

2001 – Wikipedia's foundation year – have higher edge significance, and it makes sense for Wikipedia to start adding biographies in order of importance – i.e. having a higher bar of entrance at the beginning.

There are a few considerations we must make to contextualize our results and avoid over-generalization.

First, we focus only on the 1750-1950 period: this is a somewhat arbitrary choice we take to combat record sparsity in earlier years and record abundance for contemporary people. While we think our filters are strongly motivated, we should not generalize our findings outside the 1750-1950 period without further research that can address our concerns.

Second, we have encountered the problem of some edges having zero weight, because they are present in the page, but outside the biographical text. While most of these edges indeed represent spurious relationships – as we argue in the Methods section – it is possible that some are genuine and would be valuable to keep. We need further research to discern which of these zero weighted edges are important and which are not.

Finally, here we adopt a binary perspective, distinguishing on the basis of gender between males and females. We have decided not to include in the network profiles that do not fit into this binary distinction. Information about non-binary people pre-1950 is sparse, thus this choice does not affect our results. However, if we want to move forward and include post-1950, as well as for reasons of inclusiveness, we should overcome this limitation in future works.

5. Methods

5.1. *Estimating Edge Weights*

We decide to estimate the weight of an edge by counting the number of times a person (the edge's target) is mentioned in another person's (the edge's source) Wikipedia biography. To do so, we need to perform the task of Named Entity Recognition (NER).

To solve NER, we consider two pipelines. Both have in common the use of the same resources for the ontology [44], the same algorithm to build the dependency tree [8], and WordNet as the lexical database [13]. In the first pipeline (LG) we use the explosion vectors [2], while the second pipeline (TRF) relies on transformers – specifically on RoBERTa [24]. Implementation-wise, we rely on the spaCy models (`https://spacy.io/`).

To calculate the weights we focus on the named entities that are labeled PERSON, since dealing with pronouns is difficult due to ambiguities. Moreover, pronouns would not necessarily alter the relative counts too much, because they are usually used only in proximity of the actual name. However, this introduces additional error in the weight estimation, which we deal with in the next section.

To choose between LG and TRF, we consider both computational efficiency and accuracy. LG is faster than TRF. When it comes to accuracy, by looking at

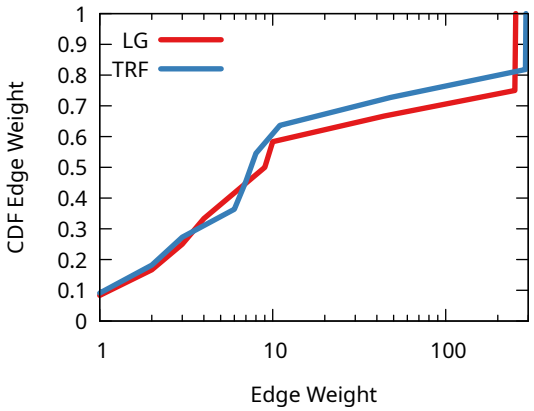

Fig. 8. The cumulative distribution (y axis) of the edge weights (x axis) from a given Named Entity Recognition method (color).

randomly sampled nodes, we find no significant difference between the weights of their edges between LG and TRF. Figure 8 shows a sample of around 1,000 edge weights – we exclude zero weighted edges in the figure because of the logarithmic axis needed for the skewed distribution.

A log-log regression between them produces an intercept of 0.009 ± 0.005, showing essential agreement on low-weighted edges, and a slope of 0.946 ± 0.007. The slope being lower than one means that LG systematically underweights compared to TRF, but the $R^2 = 0.94$ confirms that, besides this systematic underweighting, the two weight populations are linearly comparable.

Since LG and TRF arguably produced comparable outputs, we prefer the LG pipeline due to its computational efficiency.

5.2. *Edge Weight Significance*

Once we have edge weights, we can estimate the statistical significance of each of them. In practice, we assume that there might be some measurement error in the edge weights. We already talked about the issue with pronouns but, additionally, the number of mentions made in a Wikipedia page could be higher or lower than the one that should have been made. We can ask whether the number of mentions we observe is statistically different than zero, considering how many times an individual is referred to overall in the network.

This is a process known as network backboning [27]. There are a few methods to perform this task [37, 15]. The one most suitable for our setting – a directed network, with discrete edge weights, representing counts, and broadly distributed (see Figure 4) – is the noise-corrected backboning [11].

With noise-corrected backboning, the p-value tells us the likelihood of the observed edge weight to be zero. The lower the p-value, the more significant the edge

weight. The method naturally handles edge directions, as it constructs its null model with a hypergeometric Bayesian prior, estimating both the likelihood of the source node to emit a specific weight and of the target node to receive it.

The process behind the backboning procedure is an extraction without replacement model. The observed weight of the edge connecting from node u to node v is the number of successful extractions. We build a prior expectation by knowing how much weight points out from node u and how much weight is pointing towards v. We then calculate the cumulative density function of this observation, which is the probability of observing the given number of successes (edge weight) or a higher one. This is our p-value. A high value means that there is a high chance of the observed weight to be equal to or lower than the one we would expect given our null hypothesis.

In the paper, we use the p-value as the strictness criterion for the significance of the edges. The lower the p-value, the most strict we are to include an edge – that is why we reverse the x axis of Figures 5 and 6, to go from left (low strictness, high p-value) to right (high strictness, low p-value).

Appendices

S1 File. Supplementary text and figures. A document containing additional analysis to contextualize and expand over the main results of the paper.

S2 File. Data and code for reproducibility. A zip file containing the data we scraped and preprocessed from Wikipedia, as well as all the code we used to all the analyses and figures in the paper.

References

[1] Closing the gender gap: Women in red's efforts to add more women to wikipedia, `https://wikimediafoundation.org/news/2023/03/01/closing-the-gender-gap-women-in-reds-efforts-to-add-more-women-to-wikipedia/`, accessed: 2023-08-28.

[2] Explosion vectors, `https://github.com/explosion/spacy-vectors-builder`, accessed: 2023-08-28.

[3] Top websites ranking, `https://www.similarweb.com/top-websites/`, accessed: 2023-08-29.

[4] Adams, J., Brückner, H., and Naslund, C., Who counts as a notable sociologist on wikipedia? gender, race, and the "professor test", *Socius* **5** (2019) 2378023118823946.

[5] Adler, B. T., De Alfaro, L., Mola-Velasco, S. M., Rosso, P., and West, A. G., Wikipedia vandalism detection: Combining natural language, metadata, and reputation features, in *Computational Linguistics and Intelligent Text Processing: 12th International Conference, CICLing 2011, Tokyo, Japan, February 20-26, 2011. Proceedings, Part II 12* (Springer, 2011), pp. 277–288.

[6] Beytía Reyes, P. and Wagner, C., Visibility layers: a framework for systematising the gender gap in wikipedia content, *Internet Policy Review* **11** (2022) 1–22.

[7] Candia, C., Jara-Figueroa, C., Rodriguez-Sickert, C., Barabási, A.-L., and Hidalgo,

C. A., The universal decay of collective memory and attention, *Nature human behaviour* **3** (2019) 82–91.

[8] Choi, J. D., Tetreault, J., and Stent, A., It depends: Dependency parser comparison using a web-based evaluation tool, in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (2015), pp. 387–396.

[9] Clauson, K. A., Polen, H. H., Boulos, M. N. K., and Dzenowagis, J. H., Scope, completeness, and accuracy of drug information in wikipedia, *Annals of Pharmacotherapy* **42** (2008) 1814–1821.

[10] Collier, B. and Bear, J., Conflict, confidence, or criticism: An empirical examination of the gender gap in wikipedia, in *Proceedings of the ACM 2012. Conference on Computer Supported Cooperative Work, New York* (2012), pp. 383–392.

[11] Coscia, M. and Neffke, F. M., Network backboning with noisy data, in *2017 IEEE 33rd international conference on data engineering (ICDE)* (IEEE, 2017), pp. 425–436.

[12] Eom, Y.-H., Aragón, P., Laniado, D., Kaltenbrunner, A., Vigna, S., and Shepelyansky, D. L., Interactions of cultures and top people of wikipedia from ranking of 24 language editions, *PloS one* **10** (2015) e0114825.

[13] Fellbaum, C., *WordNet: An electronic lexical database* (MIT press, 1998).

[14] Giles, J., Internet encyclopaedias go head to head, *Nature* **438** (2005) 900–901.

[15] Grady, D., Thiemann, C., and Brockmann, D., Robust classification of salient links in complex networks, *Nature communications* **3** (2012) 864.

[16] Graells-Garrido, E., Lalmas, M., and Menczer, F., First women, second sex: Gender bias in wikipedia, in *Proceedings of the 26th ACM conference on hypertext & social media* (2015), pp. 165–174.

[17] Graham, M., Hogan, B., Straumann, R. K., and Medhat, A., Uneven geographies of user-generated information: Patterns of increasing informational poverty, *Annals of the Association of American Geographers* **104** (2014) 746–764.

[18] Johnson, I., Lemmerich, F., Sáez-Trumper, D., West, R., Strohmaier, M., and Zia, L., Global gender differences in wikipedia readership, in *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 15 (2021), pp. 254–265.

[19] Konieczny, P. and Klein, M., Gender gap through time and space: A journey through wikipedia biographies via the wikidata human gender indicator, *New Media & Society* **20** (2018) 4608–4633.

[20] Kumar, S., West, R., and Leskovec, J., Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes, in *Proceedings of the 25th international conference on World Wide Web* (2016), pp. 591–602.

[21] Lam, S. T. K., Uduwage, A., Dong, Z., Sen, S., Musicant, D. R., Terveen, L., and Riedl, J., Wp: clubhouse? an exploration of wikipedia's gender imbalance, in *Proceedings of the 7th international symposium on Wikis and open collaboration* (2011), pp. 1–10.

[22] Langrock, I. and González-Bailón, S., The gender divide in wikipedia: Quantifying and assessing the impact of two feminist interventions, *Journal of Communication* **72** (2022) 297–321.

[23] Laouenan, M., Bhargava, P., Eyméoud, J.-B., Gergaud, O., Plique, G., and Wasmer, E., A cross-verified database of notable people, 3500bc-2018ad, *Scientific Data* **9** (2022) 290.

[24] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V., Roberta: A robustly optimized bert pretraining approach, *arXiv preprint arXiv:1907.11692* (2019).

[25] McDowell, Z. J. and Vetter, M. A., *Wikipedia and the Representation of Reality*

(Taylor & Francis, 2022).
[26] Miquel-Ribé, M. and Laniado, D., Wikipedia culture gap: quantifying content imbalances across 40 language editions, *Frontiers in Physics* **6** (2018) 54.
[27] Neal, Z. P., backbone: An r package to extract network backbones, *PloS one* **17** (2022) e0269137.
[28] Popat, K., Mukherjee, S., Strötgen, J., and Weikum, G., Credibility assessment of textual claims on the web, in *Proceedings of the 25th ACM international on conference on information and knowledge management* (2016), pp. 2173–2178.
[29] Potthast, M., Stein, B., and Gerling, R., Automatic vandalism detection in wikipedia, in *Advances in Information Retrieval: 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings 30* (Springer, 2008), pp. 663–668.
[30] Pradel, F., Biased representation of politicians in google and wikipedia search? the joint effect of party identity, gender identity and elections, *Political Communication* **38** (2021) 447–478.
[31] Reagle, J. and Rhue, L., Gender bias in wikipedia and britannica, *International Journal of Communication* **5** (2011) 21.
[32] Rosenzweig, R., Can history be open source? wikipedia and the future of the past, *The journal of American history* **93** (2006) 117–146.
[33] Roy, D., Bhatia, S., and Jain, P., Information asymmetry in wikipedia across different languages: A statistical analysis, *Journal of the Association for Information Science and Technology* **73** (2022) 347–361.
[34] Saez-Trumper, D., Online disinformation and the role of wikipedia, *arXiv preprint arXiv:1910.12596* (2019).
[35] Samoilenko, A., Lemmerich, F., Weller, K., Zens, M., and Strohmaier, M., Analysing timelines of national histories across wikipedia editions: A comparative computational approach, in *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11 (2017), pp. 210–219.
[36] Schich, M., Song, C., Ahn, Y.-Y., Mirsky, A., Martino, M., Barabási, A.-L., and Helbing, D., A network framework of cultural history, *science* **345** (2014) 558–562.
[37] Serrano, M. Á., Boguná, M., and Vespignani, A., Extracting the multiscale backbone of complex weighted networks, *Proceedings of the national academy of sciences* **106** (2009) 6483–6488.
[38] Shachaf, P. and Hara, N., Beyond vandalism: Wikipedia trolls, *Journal of Information Science* **36** (2010) 357–370.
[39] Sun, J. and Peng, N., Men are elected, women are married: Events gender bias on wikipedia, in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (2021), pp. 350–360.
[40] Tramullas, J., Garrido-Picazo, P., and Sánchez-Casabón, A. I., Research on wikipedia vandalism: a brief literature review, in *Proceedings of the 4th Spanish Conference on Information Retrieval* (2016), pp. 1–4.
[41] Tripodi, F., Ms. categorized: Gender, notability, and inequality on wikipedia, *New Media & Society* **25** (2023) 1687–1707.
[42] Wagner, C., Garcia, D., Jadidi, M., and Strohmaier, M., It's a man's wikipedia? assessing gender inequality in an online encyclopedia, in *Proceedings of the international AAAI conference on web and social media*, Vol. 9 (2015), pp. 454–463.
[43] Wang, A., Pappu, A., and Cramer, H., Representation of music creators on wikipedia, differences in gender and genre, in *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 15 (2021), pp. 764–775.

[44] Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., Xue, N., Taylor, A., Kaufman, J., Franchini, M., *et al.*, Ontonotes release 5.0 ldc2013t19, *Linguistic Data Consortium, Philadelphia, PA* **23** (2013) 170.

[45] Wong, K., Redi, M., and Saez-Trumper, D., Wiki-reliability: A large scale dataset for content reliability on wikipedia, in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2021), pp. 2437–2442.

[46] Yasseri, T., Spoerri, A., Graham, M., and Kertész, J., The most controversial topics in wikipedia, *Global Wikipedia: International and cross-cultural issues in online collaboration* **25** (2014) 25–48.

[47] Yu, A. Z., Ronen, S., Hu, K., Lu, T., and Hidalgo, C. A., Pantheon 1.0, a manually verified dataset of globally famous biographies, *Scientific data* **3** (2016) 1–16.