Wiki Workshop (10th edition) – May 11, 2023

Gender Asymmetries in the depiction of Historical Figures: A Comparison of Bede's Historia Ecclesiastica and Wikipedia

Yurong Wang
RWTH Aachen University
Aachen, Germany

Claudia Wagner
GESIS
Cologne, Germany
RWTH Aachen University
Aachen, Germany

Ana L. C. Bazzan
UFRGS
Porto Alegre, Brazil

Abstract

This study compares the gender portrayal of historical figures in Wikipedia with the early medieval historical work by Bede. The investigation examines how individuals in Bede's work are selected, presented, and placed within the network on Wikipedia. The findings reveal significant gender biases, with women historical figures being less likely to be considered notable, missing crucial biographical information, and ranked lower in the hyperlink network. These biases highlight the need for more efforts to address gender disparities on Wikipedia.

Keywords: Gender bias, Wikipedia, Wikidata, Network structure, History characters

Introduction

The representation of historical figures is crucial for shaping our understanding of the past and present. Gender bias is not limited to traditional forms of media and records, but also prevalent in digital media and online platforms, including Wikipedia (Wagner et al., 2016). This work aims to contribute to the existing scholarship on gender asymmetries by comparing the representation of historical figures in Wikipedia to Bede's Historia ecclesiastica gentis Anglorum (Ecclesiastical History of the English People, HE). By analyzing the depiction of the same group of historical figures in both resources, it aims to provide insights into how gender asymmetries are reflected in the portrayal of historical figures on Wikipedia and how it differs from Bede's original text from the early medieval period.

Methods

The HE character list [1] provided information on 594 characters, including their names, gender, and appearances in specific chapters. Each character was then mapped to a corresponding Wikidata entity and their Wikipedia pages. The Wikidata SPARQL query service (WSQS) [2] was used to obtain the number of language editions for each character, a list of notability identifiers used to quantify each character's level of notability, as well as various biographical and relational properties. Hyperlinks on Wikipedia pages were obtained using the SPARQL endpoint over the DBpedia dataset[3].

The first research question (RQ1) investigates whether male and female historical figures are selected for Wikipedia pages based on the same criteria. Two measures are employed as proxies of notability: the count of language editions and the count of associated identifiers that suggest notability. Considering the excess zeros in both counts, zero-inflated negative binomial (ZINB) regression is used to analyze the factors that potentially influence notability counts for male and female historical characters, with occupation[4] as a confounding variable.

The second research question (RQ2) investigates whether there is a significant association between the gender and the type of available information. Chi-squared tests are applied to compare the observed frequencies of metadata properties for male and female historical characters to the expected frequencies.

Finally, the third research question (RQ3) examines whether male and female historical characters have different structural properties in the Wikipedia link network. The network of historical characters' Wikipedia pages is constructed using hyperlinks on Wikipedia pages. Communicability, PageRank, and betweenness are used as centrality measures to rank all figures. While the size of the top-ranked window increases, the proportion of women is tracked and then compared to 3 types of constructed graphs (random, graph preserving the degree sequence, and small world graph). Besides centrality measures, entropy variation of communicability values upon the characters' removal are used to capture the importance of a character on the network topology - nodes causing the larger entropy change are considered to be more influential. This metric is also utilized to distinguish females with more *independent* importance, regardless of the impact of their male kinships. (Prado et al., 2020)

[1] The character list was compiled by J. Hillner, M. MacCarron, U. Vihervalli, and R. Heffron, and licensed under CC BY-NC 4.0.

[2] See Wikidata Query Service.

[3] See SPARQL Query.

[4] Occupation information used for ZINB model is adopted from the level-2-main-occupation in the publicly available cross verified dataset (Laouenan et al., 2022).

Results

RQ1 first looked at the group that didn't make it into Wikidata, and found a higher proportion of unnamed female characters than males. The former were often described in relation to family or marital status (e.g., "Wife of Sebbi"), while the latter were linked more to occupation (e.g., "Aged priest"). The results from ZINB models suggest that female historical figures are less likely to have a Wikipedia page, but more likely to have multiple language editions and notable identifiers, indicating their higher notability in the world of Wikipedia.

RQ2 uncovers potential gender biases in the representation of historical figures on Wikipedia during the content building stage. The findings suggest that certain metadata properties, such as occupation and religion, are better covered for male historical figures, while familial relation properties, such as spouse, child, father, and mother, are better covered for female historical figures. Chi-square tests show a highly significant association between gender and the availability of properties, while women may be more likely to have their kinship relationships emphasized over their occupational achievements.

RQ3 looked at the gender inequalities in the structural placement in the graph created using Wikipedia hyperlinks (Graph-Wiki) including 2,257 edges and 387 nodes. The author computed communicability values for each character and identified the top-tier characters based on the entropy variations of the network upon their individual removal. The top 13 most impacting characters on Graph-wiki's topology are all male, while 2 out of 13 are female in Graph-HE. After ranking all characters based on 3 centrality measures separately, the author compared the results from Graph-Wiki with three baseline models (Random, Degree sequence, and Small world). Results in Figure 1 indicate that the women fraction in top-n characters in random and small world graphs converge to the expected women fraction rapidly with all three centrality measures, while the empirical graph and the degree sequence-preserving graph do not converge till the very end; women consistently exhibit lower centrality in the top 50-100 nodes of the empirical graph than in the degree sequence-preserving graph, suggesting that factors beyond network structure heterogeneities may contribute to their limited centrality. Figure 1 (a) shows a short surge in the proportion of women in the top-50 nodes, indicating specific women in the graph have high influence despite the overall under-representation of women. As communicability centrality measures the communication flow through nodes, taking into account all possible paths with a penalty for longer paths, some nodes may have high communicability despite being connected to only a limited number of nodes with central roles. These nodes may have limited direct connections to the broader network but can still spread information and influence others through powerful channels around them. This pattern may suggest a form of marginalization where women face structural barriers limiting their direct access to broader networks.

Discussion/Conclusions

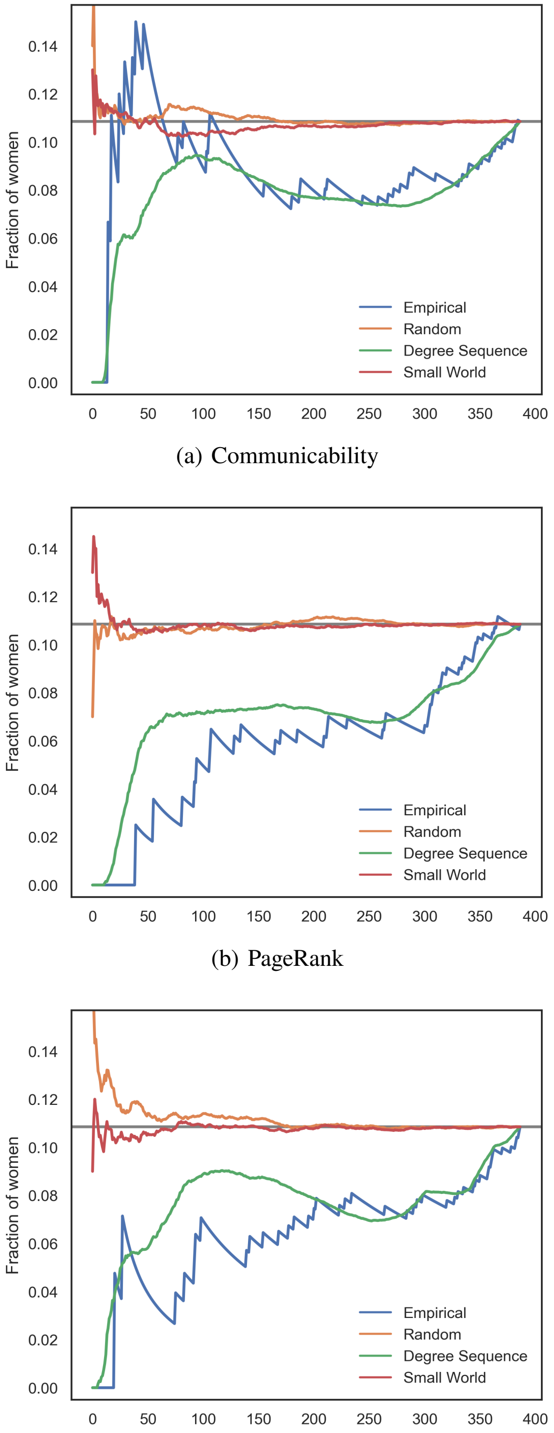
The analysis on Wikipedia's selection process exposes gender asymmetries in this process, where women characters may face stricter criteria to be deemed notable enough for a Wikipedia page, resulting that female historical figures are more likely to be excluded from Wikipedia, and they may face additional barriers compared to their male counterparts to get included in Wikipedia. The study also reveals gender disparities in the completeness of biographical information available on Wikipedia, with women historical figures being less advantaged in certain areas, such as date of death, occupation, and religion. This systemic bias in information presented on Wikipedia can unintentionally amplify existing gender biases in historical records and result in unpredictable and far-reaching effects. The hyperlink network shows a different form of gender bias, with female historical figures being relatively marginalized in the network structure. Although some female historical figures have links to important nodes and have the potential to achieve high influence in the network information flow, they still face structural barriers that limit their direct connections to broader network. This marginalization of female historical figures in the network structure may have implications for their visibility and recognition. In conclusion, the study sheds light on the gender disparities in the selection process of historical figures for inclusion in Wikipedia, the completeness of biographical information available on Wikipedia, and shows the marginalization of female historical figures in the network structure. These findings suggest the need for increased efforts to address systemic biases in the depiction of historical figures on Wikipedia.

References

[Laouenan et al.2022] Morgane Laouenan, Palaash Bhargava, Jean-Benoît Eyméoud, Olivier Gergaud, Guillaume Plique, and Etienne Wasmer. 2022. A cross-verified database of notable people, 3500BC-2018AD. *Scientific Data*, 9(1):290, December.

[Prado et al.2020] S. D. Prado, S. R. Dahmen, A. L. C. Bazzan, M. Maccarron, and J. Hillner. 2020. Gendered networks and communicability in Medieval Historical Narratives. *Advances in Complex Systems*, 23(03):2050006, May.

[Wagner et al.2016] Claudia Wagner, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. 2016. Women through the glass ceiling: gender asymmetries in Wikipedia. *EPJ Data Science*, 5(1):5, December.



(a) Communicability

(b) PageRank

(c) Betweenness

Figure 1: Fraction of women in top k characters ranked by 3 centrality measures; The grey line indicates the expected women fraction, which corresponds to the actual fraction of women in all characters in Graph-Wiki.