OSF Preregistration: Wikilink Analysis

Study Information

1. Title

Feminist Interventions on Wikipedia: From Collaboration to Content

2. Authorship

Isabelle Langrock[1]

[1] University of Pennsylvania

3. Description

With over 5 million entries in the English language edition alone, Wikipedia is the world's largest collaborative endeavor. However, equally well documented is Wikipedia's gender problem: not only do women participate as editors at significantly lower levels than men, even when compared with gender gaps in other technological pursuits, but biographies of women are often incomplete, poorly edited, or even remain unwritten. Scholars from STS, Communication, Media Studies, and Computational Social Science have examined the persistence and vastness of these gaps and even made suggestions for improvement, yet few have examined the effects of the initiatives attempting to remedy the disparity. This project examines how feminist movements reconceptulize collaboration as a process of sharing values and motivations rather than a system that allows many people to participate, as described in Wikipedia's norms. Establishing these two forms of collaboration, this project also looks at how they impact the structure of the content that is produced by examing Wikipedia's algorithmic assessment of article quality and the wikilink structure, both of which have impacts beyond the site, between female biographical articles edited by the movements and other groups that have no interventions: female politicans, athletes, and academics. The data includes a corpus of training materials and mission statements from the two movements and Wikipedia for qualitative thematic analysis of how collaboration is thought about by all three populations. Additional data for computational assesment includes the biographical articles and their revision history from the movement's dashboards tracking edits as well as other identified biographical articles for comparison.

This research project serves as an example of the combined socio-techinical effects of any digital collaborative endevor on the product. Furthermore, the focus on feminist interventions on a platform with a prevalent gender gap refocuses attention away from a deficeit based model that looks at what women are not doing and towards a framework that suggests how sociotechnical processes can be hostile, and become more open, to marginalizes populations.

4. Hypotheses

1. **Wikilinks**
   Do the articles edited by the feminist movements (500 Women Scientists and Art+Feminism) link to each other more than other biographical articles grouped by occupation and more than to be expected by random chance?

Given the motivations of collaboration of the two groups to make their histories visible and placing emphasis on the value of gathering together to edit Wikipedia and learn from each other, I expect to find a more dense network within the scientist and artist groups than the other groups.

Design Plan

5. Study Type

- Observational Study - Data is collected from study subjects that are not randomly assigned to a treatment. This includes surveys, natural experiments, and regression discontinuity designs.

6. Blinding

- No blinding is involved in this study.

7. Is there any additional blinding in this study?

There are no blinding procedures in this study.

8. Study design

This study will build a citation network of nodes with nodes as the biographical articles and edges representing the wikilinks from one article to another. Network density and modularity will be compared across the two "treatment" groups of articles pulled from the movement's edits dashboards (see below) and groups of other biographies sorted by profession.

9. Randomization

Randomization is not applicable.

Sampling Plan

The sample for this research project is pulled from the recorded edits on the two movement's dashboards, a tool provided by Wikipedia for groups to track their edits, then reduced to only the biographical content. All articles edited that are not biographical (i.e. about theories, concepts, objects, and/or places) are ommitted from the sample. Groups of similiar size are selected for the "Control" groups using the names included on Wikipedia lists for each profession/group. Lists were collected by searching Wikipedia for "Female X profession" and then selecting names from each list available. Duplicates were removed.

10. Existing data

- Registration prior to analysis of the data: As of the date of submission, the data exist and you have accessed it, though no analysis has been conducted related to the research plan (including calculation of summary statistics). A common situation for this scenario when a large dataset exists that is used for many different studies over time, or when a data set is randomly split into a sample for exploratory analyses, and the other section of data is reserved for later confirmatory data analysis.

11. Explanation of existing data

The data used in this study is publically available as it is public commons content from Wikipedia. Additionally, research conducted on the movements has led to a familiarity with Wikipedia edits and individual biogrpahies. However I have not analyzed the large data set nor have I assembled the data for the "control" conditions prior to making the hypotheses.

12. Data collection procedures

Data is collected by assembling a list of biographical pages from the movements' edit dashboards and Wikipedia lists of "Women in X" where x is another proffesional/occuptaional category, matching the movements' respective emphasises on scientists and artists. Data about each article will be then collected using the Wikipedia API, including a list of wikilinks for each article.

13. Sample size

Sample size is determined by the number of biographical pages Art+Feminsim and 500 Women Scientists have claimed to edited through the WikiMedia dashboard tracking tool. The other professional groups was gathered through the lists. Duplicates and nonbiographical information is then removed.

14. Sample size rationale

Sample size was determined by how many pages the movement's have edited.

15. Stopping rule

Not applicable.

Variables

16. Manipulated variables

Not applicable.

17. Measured variables

- Network density: measuring how well connected (dense) each network, seperated by occupation, is. This is a ratio of the number of edges observed and the number of edges possible in the network.
- Network modularity: modularity is a form of community detection, detailing how many groups (of nodes more highly connected to each other than others in the network) are in the network. While density tells us about the whole network, modularity is able to give us a better picture of how the network is dispersed.

18. Indices

Not applicable

Analysis Plan

I will build a network of the wikilinks for each group, where: - nodes = biographical figure/subject of the article - edges = wikilinks between articles; direction of edge showing what page the link is from.

Then I will analyze the density and modularity of each network and compare them to each other and some random networks that preserve key features of the original networks (number of nodes and edges, etc)

19. Statistical models

Describe your planned statistical model(s) here.

20. Transformations

Not applicable.

21. Inference criteria

Describe your inference criteria here or state not applicable.

22. Data exclusion

Observations will be exluded if they: 1) Are not a biographical Wikipedia article (ex: objects, concepts, places and lists are frequently the subject of Wikipedia articles and could have been edited by the movements. However the primary focus of the movements are editing biographical pages of women so they will not be examined here) 2) I'm only looking at wikilinks to other biographies in my dataset, for example, if Frida Kahlo's Wikipedia article (which is quite robust already) is not edited by the Art+Feminism, I won't look at Wikilinks to her.

23. Missing data

Not applicable.

24. Exploratory analysis

Enter any plans for exploratory data analysis here or state not applicable.

Other

25. Other

Enter any additional information not covered by other sections, or state not applicable.

References

Enter any references used throughout the text here.