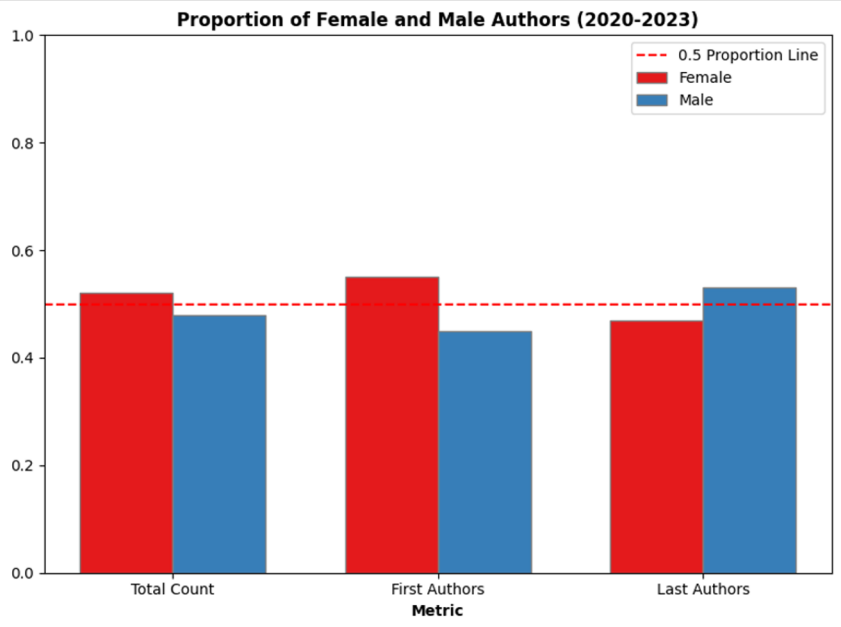
Analysis of scientific publishing data

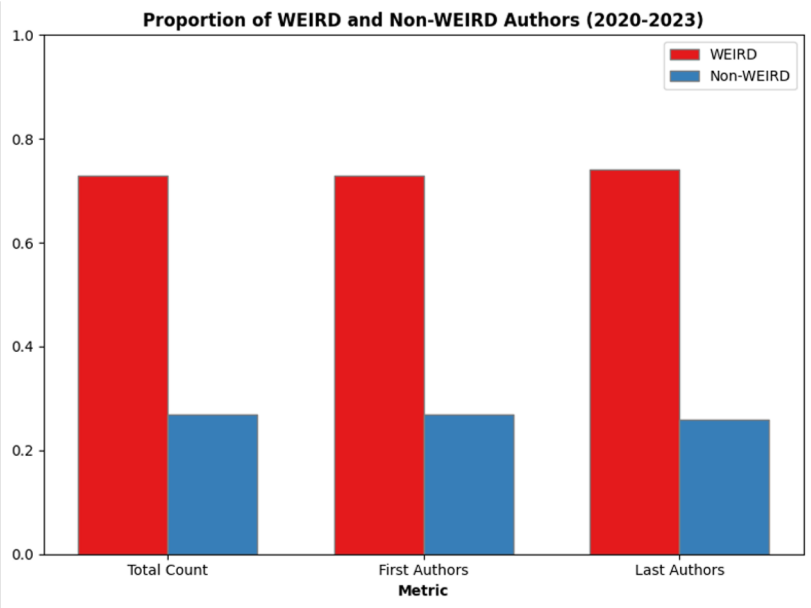
project proposal for students of Network Science 2025/2026

Lejla Džanko (ldzanko@st.swps.edu.pl)

Project on gender and geographical diversity in publishing

- Are women and authors from non-WEIRD (Western, European, Industrialized, Rich, Democratic) countries underrepresented in psychology/computer science/medicine?





Datasets explained

- metadata of articles published in psychological/medical/computer science journals in the 2020-2023 timeframe
- article, journal and author-level metadata

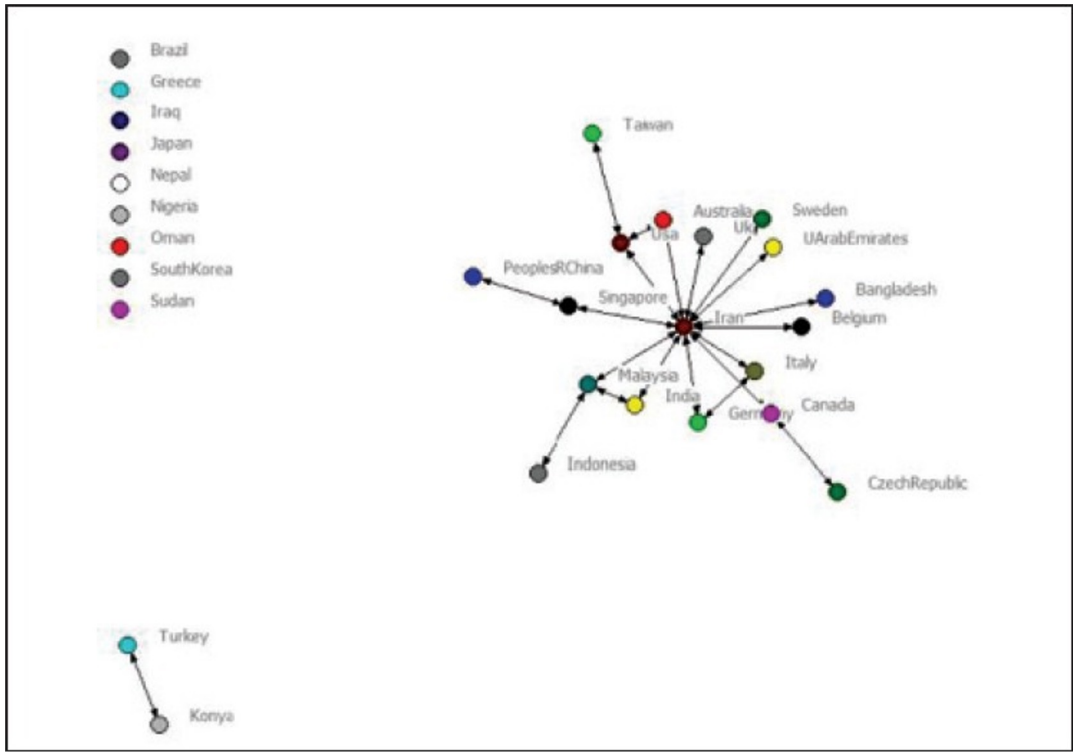
Scopus

Datasets explained

- article information: title, DOI, keywords, abstract, authors and their affiliation, year of publishing
- author information: name, institution, country, gender
- journal information: name, E/ISSN, journal impact metrics

Project idea #1 - authorship network

- connect authors who have published together (appeared on the same paper)



Project idea #1 - authorship network

Identify clusters (communities) of authors in the network and try to see:

1. Do authors collaborate more with people from their own countries or abroad? Visualize the network on the world map.
2. Do WEIRD and non-WEIRD authors collaborate? If so, who are the authors that are bridges between them and why (requires exploration of their profile)?
3. How do authors also cluster together based on the impact metrics of journals they publish in? If there is a low/high impact metric cluster, what is the gender and geographical composition of the clusters?

Project idea #2

Analyze keywords of women-only vs male-only articles

1. Create network of keywords that appear together. Cluster them and check for themes/topics that appear. How likely is that the keyword appears in a male vs. female article (calculate gender bias of the word)?
2. Assign scores (agency and communion, among others) to keywords. Create networks based on score correlations. Check keywords cluster for gender composition. Are male (vs. female) keywords more likely to be e.g. agentic (vs. communal)?

Same can be done for WEIRD vs. non-WEIRD articles and also using abstracts/words from abstracts instead of keywords.

Pros

- project idea fully formulated and feasible (but of course input is welcome)
- datasets are already downloaded and formatted
- I will act as a team member (but not do the work for you)
- real-life, interesting data
- currently very popular topic
- novel work
- possibility of a publication if project fully developed

Non-pros

- large datasets that require careful handling (sampling etc.)
- some reading of the literature to understand concepts could be helpful
- a topic that is a bit more involved
- ideally you only do it if you are really interested in the project and not just the grade

Thank you for your attention