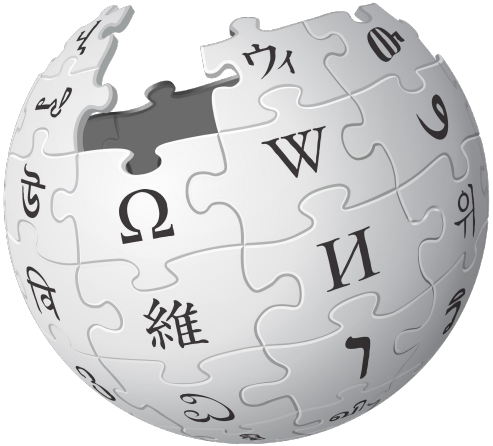
Exploring Network Dynamics and Centrality within the Wiki-Topcats Network

Caleb Miller and Luke Chesley
INFO 623 - 001 - Term Project

Wikipedia’s Representation of Human Knowledge

- Wikipedia contains an immense network of articles connected through hyperlinks
  - This presents a unique opportunity to explore the structure of human knowledge
- The analysis of this network will uncover patterns that reveal:
  - How information is interconnected
  - Which topics are deemed most important
  - How knowledge communities cluster together.
- Goal of this study:
  - We aim to thoroughly analyze the Wikipedia hyperlink graph to answer questions about the network’s structure and what it tells us about human knowledge as a whole.



Related Work

- Previous studies have analyzed Wikipedia's hyperlink graph to understand article quality and editor behavior, revealing insights into the network's structure and potential biases.
- These analyses show that top quality articles have higher in and out-degrees
    - Suggesting broad influence within the wikigraph
- Research on the graph and editing patterns indicates potential biases in how information is connected and presented
    - This is driven by both the structure of Wikipedia and the biases of its editors

Research Questions

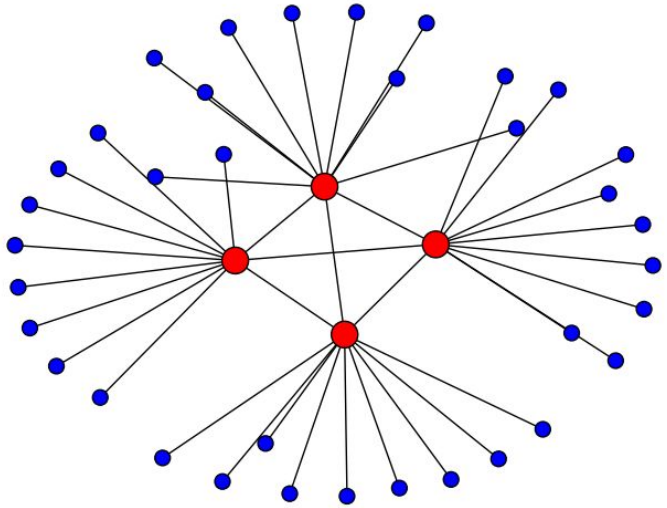
- Centrality and Prestige: Which Wikipedia articles are most central in the graph, indicating high prestige?
- Influence and Authority: Which pages have the most authority according to algorithms like PageRank, and how does this compare to their centrality?
- What does the network's structure reveal about the connectivity of represented countries?
- Our hypothesis:
    - Centrality metrics and algorithms such as PageRank can serve as indicators of article importance
    - The shape of the network will show clear patterns of connectivity that relate to geography and topic

Data Collection

- Primary dataset collected from the Stanford Large Network Dataset Collection
  - “Wikipedia Top Categories” dataset - hyperlink dataset of Wikipedia
  - Available at: https://snap.stanford.edu/data/wiki-topcats.html
- Dataset provides an all-encompassing look at Wikipedia’s hyperlink structure
  - Offers a rich foundation to our analysis
- Data is a snapshot of the network from September 2011
  - Restricted to pages belonging to categories with at least 100 pages
- Despite being outdated and slightly pruned, this dataset will still be able to provide insights to large structural patterns

Network Construction

- The Wiki-Topcats dataset contains nodes representing Wikipedia articles
- Edges in this dataset represent hyperlinks between articles
- For network construction, we used the ‘networkx’ library in Python
    - We created a directed network, capturing the directional nature of hyperlinks



Initial Network Analysis

- Network size: 1.79 million nodes and 28.5 million edges
- The average nodal degree of 31.83 implies a fair level of connectivity
- Diameter: 9
- 90-percentile effective diameter: 3.8
- However, a very low network density and centrality metrics tell a deeper story
    - Network Density: 8.883757461873000e-6
    - Centrality Metrics:
    - Eigenvector Centrality: 6.42805733875925e-05
    - In-Degree Centrality: 8.883757461872467e-06
    - Out-Degree Centrality: 8.883757461872738e-06
    - PageRank Centrality: 5.581948870464576e-07

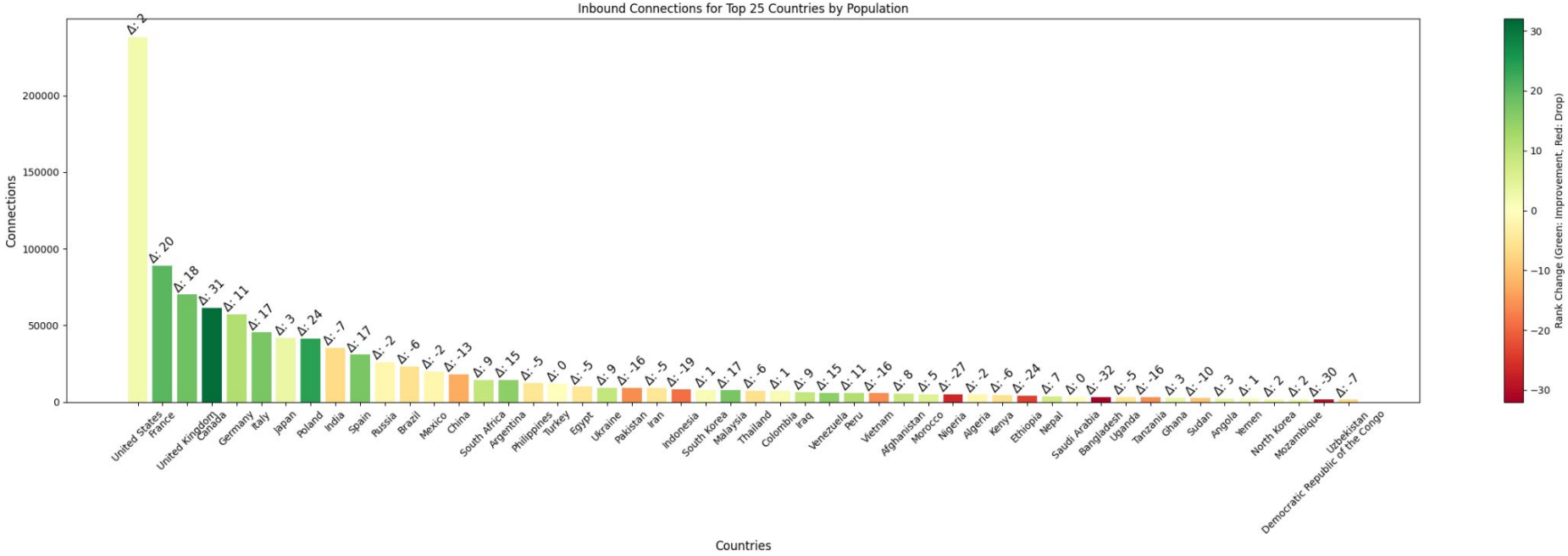
Centrality Metrics

- Initially, the low average values across centrality metrics suggested a sparse and loosely connected network
  - Typical for large networks
- This broad view ignores the significant roles played by individual nodes
  - Specifically, countries emerged as central figures in Wikipedia’s network
- Leaders for centrality metrics:
  - Eigenvector: United States, World War II, and the United Kingdom
  - In-Degree: United States, France, and the United Kingdom
  - Out-Degree: List of American Television Programs and List of English Writers
  - PageRank: United States, France, and Departments of France
- Pattern emerges with leaders surrounding the US, France, and Western ideologies

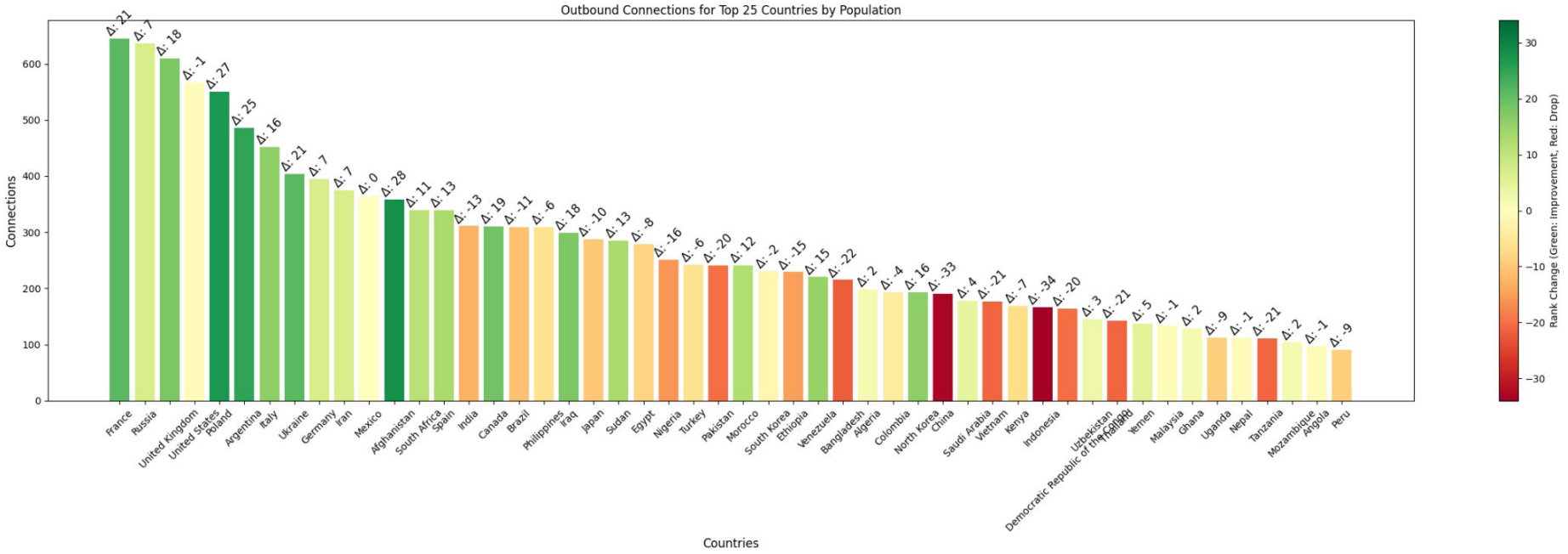
Global Connectivity Bias

- This pattern prompted an investigation on the links associated with articles of the world's most populous countries
  - We ranked these countries by population and analyzed their Wikipedia articles' connectivity
    - Notable overrepresentation of Western countries and underrepresentation of non-Western nations
  - Used the proportion of global population to estimate expected proportion of connections for each country
    - Few countries align with the assumption that population is proportional to connectivity
- Western Countries are disproportionately overconnected
  - Both in inbound and outbound links
- Large discrepancy in connectivity for China, India, and other non-Western nations

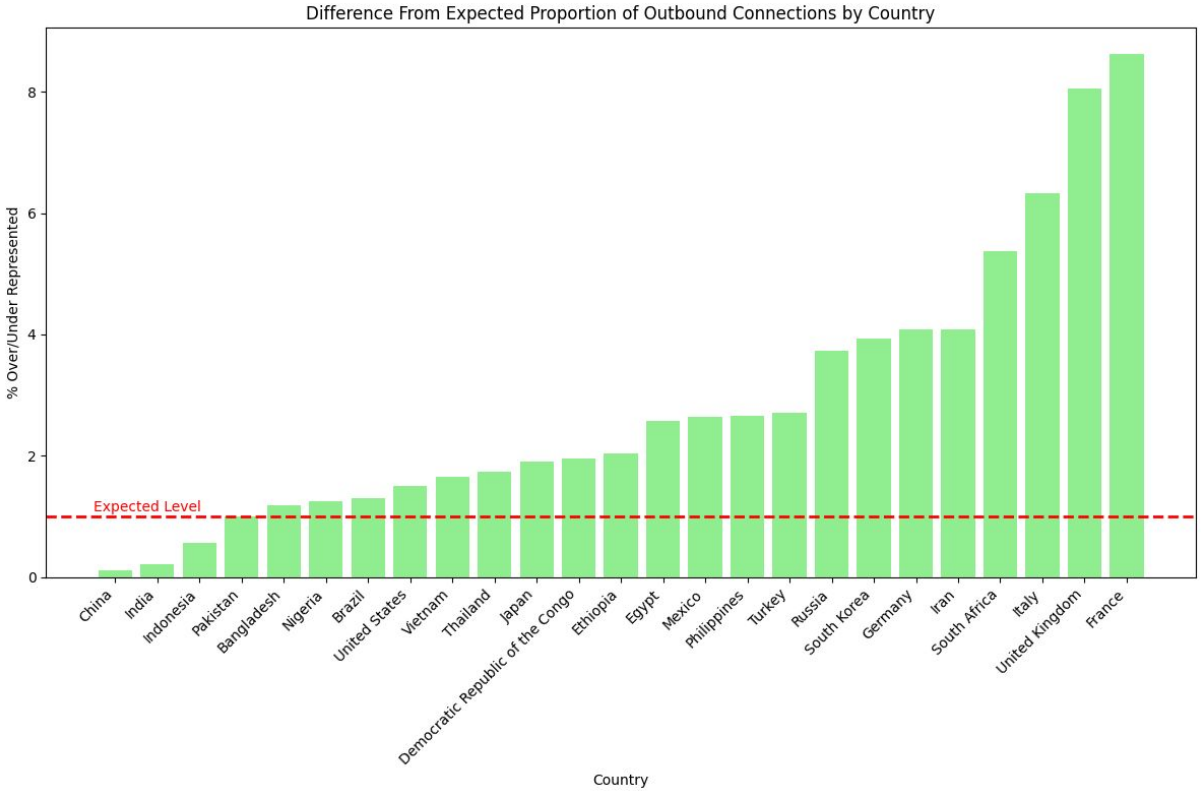
Country Population vs In-Degree Ranking



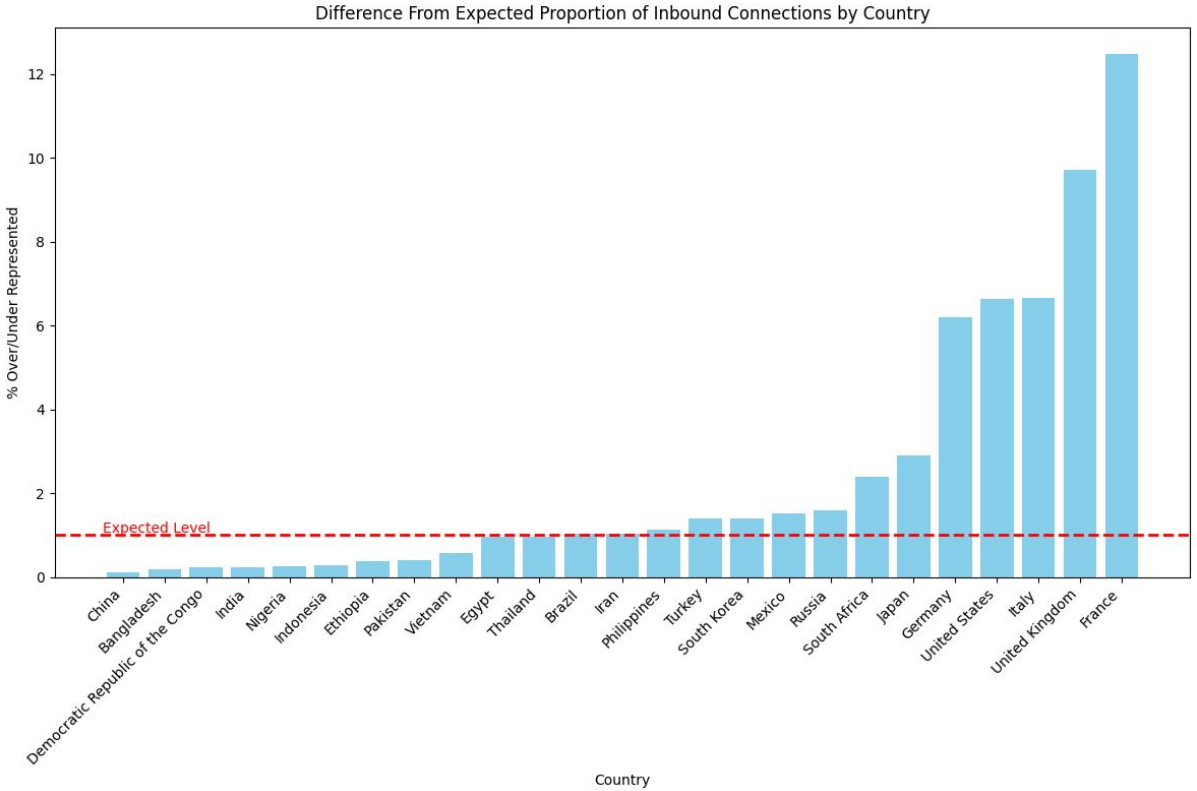
Country Population vs Out-Degree Ranking



Countries vs Expected Proportion of Outbound Connections

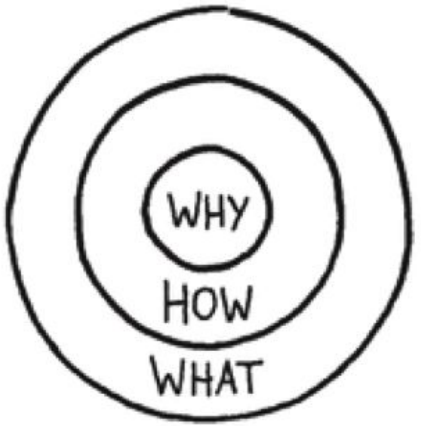


Countries vs Expected Proportion of Inbound Connections



Possible Explanations for Disparity

- Disparity in Wikipedia connectivity may be due to English serving as a *lingua franca*, influencing the overrepresentation of English and Western articles
- Our analysis was limited to the English version of Wikipedia
    - This is the most extensive and active, but does not represent the entire ecosystem
    - 6 / 10 most active Wikipedias are in Western Languages
        - Indicative of preference for Western topics



Future Directions

- Analysis of Wikipedia in other languages is needed to understand if linguistic / cultural proximity affects clustering
- Understanding of disparities is incomplete without examining patterns across different language versions of Wikipedia.
- Future research should explore the relationship between language, culture, and knowledge organization on Wikipedia