Caleb Miller and Luke Chesley
INFO 623 - 001
Term Project

Exploring Network Dynamics and Centrality in the Wiki-Topcats Dataset

Abstract:

This paper investigates the structural makeup and centrality measures of the Wiki-Topcats network. This comprehensive dataset was collected from Wikipedia's category system. The immense network of Wikipedia articles, connected through hyperlinks, presents a unique opportunity to explore the structure of human knowledge. By analyzing this network, we can uncover patterns that reveal how information is interconnected, which topics are deemed most important, and how knowledge communities cluster together. The goal of this project is 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:

Wikipedia is the most comprehensive collection of human knowledge available to the median internet user. It contains millions of articles in hundreds of languages[1], covering topics from the highly specialized to the everyday. The references between articles make up a complex web that facilitates user navigation through wikipedia but also mirrors the web-like nature of knowledge itself. The community built and ever changing structure serves as a starting point for exploring the realm of knowledge organization and uncovering potential biases on how it is distributed.There have been many studies examining the structure of the Wikipedia hyperlink graph, commonly referred to as the wikigraph, to be the basis of further analysis. Chhabra, et al.[2] used the structure of the wikigraph along with quality grades to predict quality grades of non-wikipedia articles. They used in-degree and out-degree, betweenness centrality, PageRank and other metrics to weight quality grades in their predictive model. They argue " Top quality articles are comprehensive and broad in their coverage as compared to the articles belonging to low-quality classes. Therefore, in the Wikigraph, top quality articles are expected to have more in-degree as well as out-degree."(Chhabra et al., 2020, p. 586). Higher quality articles are expected to have a higher in-degree and out-degree. Brandes, et al.[3] analyze the network of edits and editors using similar and extended methods to uncover patterns in the edits and distinct groups of editors with opposing opinions which raises discussion about potential bias in the structure wikipedia, stemming from bias in the editors. These studies led us toward uncovering biases in the distribution of nodal connections compared to expected levels, discussed in subsequent sections.

Research Question:

This study is guided by a series of research questions focused on understanding the architecture of the Wikipedia graph:

- 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?

We hypothesize that centrality metrics and algorithms such as PageRank can serve as indicators of article importance and that the shape of the network will show clear patterns of connectivity that relate to geography and topic.

Data Collection:

The primary dataset for this project was collected from the Stanford Large Network Dataset Collection, specifically the "Wikipedia Top Categories" dataset (available at https://snap.stanford.edu/data/wiki-topcats.html) . This dataset provides an all-encompassing look at Wikipedia's hyperlink structure, offering a rich foundation to our analysis. The dataset is a snapshot of the structure collected in September 2011, restricted to pages belonging to categories with at least 100 pages. Although being outdated and slightly pruned, this dataset will still be able to provide insights into large structural patterns.

Network Construction, Centrality, and Connectivity:

The Wiki-Topcats dataset contains nodes representing Wikipedia articles categorized under various topics. Edges in this dataset represent hyperlinks between articles. Using the 'networkx' library within Python, we constructed a directed network, capturing the directional nature of hyperlinks. The network is quite large and has over 1.79 million nodes and 28.5 million edges. The average nodal degree of 31.83 implies a fair level of connectivity. However, a very low network density and centrality metrics tell a deeper story about these connections and the spread of influence within the network.

- **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

Initially, the low average values across eigenvector, in-degree, out-degree, and PageRank centralities suggested a sparse and loosely connected network. This is typical for large networks where the potential for connections exceeds the actual connections made. However, this broad view ignores the significant roles played by individual nodes, particularly those representing countries, which emerge as central figures within the Wikipedia network. A closer analysis of articles with the highest centrality metrics shows a different perspective. This approach allows this study to pivot the discussion towards the influence of certain central nodes within the network.

Eigenvector Centrality leaders include the United States, World War II, and the United Kingdom. These articles demonstrate a higher level of influence because of their connectivity to other influential nodes. This suggests a network subsection where certain topics, especially countries and significant historical events, form densely interconnected clusters. This gives these nodes considerable influence.

Similarly, In-Degree Centrality leaders include articles like the United States and France as key informational destinations. These nodes attract a large number of connections, proving their importance within the network's structure as primary sources of information or reference points. Out-Degree and PageRank Centrality metrics further reinforce this pattern with leaders surrounding the United States, France, and other knowledge leaders. These articles stand out for their ability to spread information widely and are recognized as authoritative nodes.

Focusing on countries and pivotal historical events that lead the pack in centrality metrics shows a pattern of connectivity and influence. This selective density within the network contrasts the overall sparse network structure, and proves the importance of specific nodes in facilitating the flow of information in the network.

Global Connectivity Bias:

A majority of the central nodes in our analysis, regardless of the metric used, correspond to articles about countries. This pattern prompted an investigation on the links associated with articles of the world's most populous countries as of 2012, aligning with our data collection period. We ranked these countries by population and analyzed their Wikipedia articles' connectivity—both in terms of in-degree (incoming connections) and out-degree (outgoing connections).
Our analysis revealed a pronounced trend, shown in Figure 1: the article on the United States stands out with the highest number of inbound links, closely followed by an unexpectedly high-ranking France—surpassing its population-based expectations by 20

positions. This discrepancy points to a notable overrepresentation compared to its global population rank, a trend similarly observed in other Western countries like Canada, Poland, Germany, and the UK. Conversely, countries such as China and India show a marked underrepresentation, indicative of a potential Western bias within Wikipedia's hyperlink structure.
The disparity becomes even starker when examining out-degree connections. France, again, leads in out-degree connections, significantly surpassing its population rank. China's articles, in contrast, fall 32 positions below what their population rank would suggest, further highlighting the imbalance.

These findings prompted further questions. We used the proportion of global population for each country as an expected proportion of in and out connections for the corresponding article. The findings are presented in Figures 3 and 4. The assumption that population is proportional to in and out degree holds true for only a small number of countries, with no obvious pattern for why it might hold true. Generally, Western countries are still proportionally over connected. The out degree results, figure 3, show that a large portion of the discrepancy in connection comes from China and India. The wide disparity in their degrees allows almost all other countries to be proportionally over connected. The in degree results, presented in figure 4, show that many more countries are proportionally connected. However, western countries are still proportionally over connected and highly populated Asian countries are still proportionally under connected.

These findings underscore a disproportionate emphasis on Western countries in Wikipedia's interconnected landscape, raising questions about the representation and visibility of non-Western nations within this global knowledge repository.

Possible explanations for disparity

One possibility for the scale of disparity shown in the analysis is that English serves as a lingua franca, or a language used as a medium of communication between people who do not share a native language. English’s prominence in global communications may naturally lead to the disproportion in the connectivity of English and western related articles. This analysis was confined to the English version of Wikipedia. English Wikipedia is by far the most extensive and active but still only represents a portion of the entire Wikipedia ecosystem. Despite this shortcoming in the analysis it is still notable that six of the ten most active Wikipedias as in other Western languages[4], suggesting other factors that favor western topics.

The disproportionate representation of Western topics prompts further analysis into the structure of the other language Wikipedias. Such analysis might reveal whether linguistic and cultural proximity influences the organization of knowledge, leading to a

clustering of content around culturally or linguistically similar countries. Without examining these patterns across different language versions of Wikipedia, our understanding of these disparities remains incomplete. Future research should aim to dissect these dynamics, offering insights into the complex interplay between language, culture, and knowledge organization on a platform as globally integrated as Wikipedia

Conclusion:

In this study, we explored the Wiki-Topcats dataset to understand the structure and centrality dynamics within Wikipedia's vast network of articles. Through this analysis, we uncovered several key insights that shed light on the organization of human knowledge as represented in this digital encyclopedia. Our investigation of centrality metrics revealed that certain nodes, particularly those representing countries and significant historical events, hold notable influence within the network. These nodes form densely interconnected clusters that are key in the spread of information.

Moreover, the analysis of global connectivity bias has brought to light a significant overrepresentation of Western countries, which contrasts with the underrepresentation of non-Western nations such as China and India. This gap implies a strong Western bias in the structure of Wikipedia's English version, which may be due to the predominance of English as a *lingua franca* and the concentration of active Wikipedia editions in Western languages. These findings show the importance of language and cultural based factors in the organization and accessibility of knowledge on Wikipedia.

This study also points out the limitations of our analysis, including our focus on the English version of Wikipedia, which only represents a fraction of the global Wikipedia network. This limitation emphasizes the need for future research on the structure of Wikipedia across different language versions to gain an all-inclusive view of knowledge organization and biases on the platform.

To conclude, our study contributes to the ongoing discussion about the structure of knowledge in the Wikipedia network and the factors that influence its organization. By diving into the centrality of certain topics and the biases in the connectivity, we reinforce the need for a more inclusive representation of human knowledge that takes account of the diversity of different cultural values and languages. Future research on this topic is necessary to address this inconsistency and work towards a more balanced and all-encompassing network of knowledge within Wikipedia.

Citations

[1] Wikipedia. Wikipedia:Size_of_Wikipedia from wikipedia, the free encyclopedia. [Online; accessed 2024-03-02].

[2] Anamika Chhabra, Shubham Srivastava, S. R. S. Iyengar, and Poonam Saini. 2021. Structural Analysis of Wikigraph to Investigate Quality Grades of Wikipedia Articles. In Companion Proceedings of the Web Conference 2021 (WWW '21). Association for Computing Machinery, New York, NY, USA, 584–590. https://doi.org/10.1145/3442442.3452345

[3] Ulrik Brandes, Patrick Kenis, Jürgen Lerner, and Denise van Raaij. 2009. Network analysis of collaboration structure in Wikipedia. In Proceedings of the 18th international conference on World wide web (WWW '09). Association for Computing Machinery, New York, NY, USA, 731–740. https://doi.org/10.1145/1526709.1526808

[4] Wikipedia. Wikipedia:List_of_Wikipedias from wikipedia, the free encyclopedia. [Online; accessed 2024-03-02].

Caleb Miller and Luke Chesley
INFO 623 - 001
Term Project

Figures:

Fig 1. - In-degree of top 25 most populous countries, colored and annotated by the difference from their population ranking.

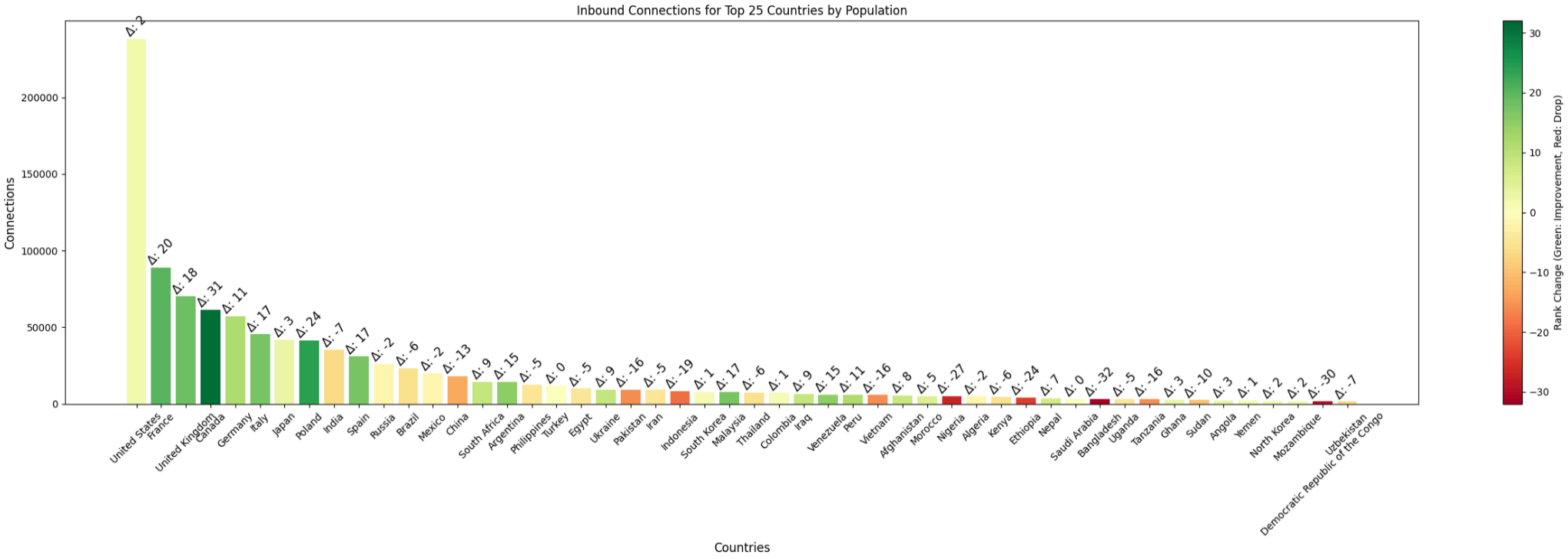

Fig 2. - Out-degree of top 25 most populous countries, colored and annotated by the difference from their population ranking.

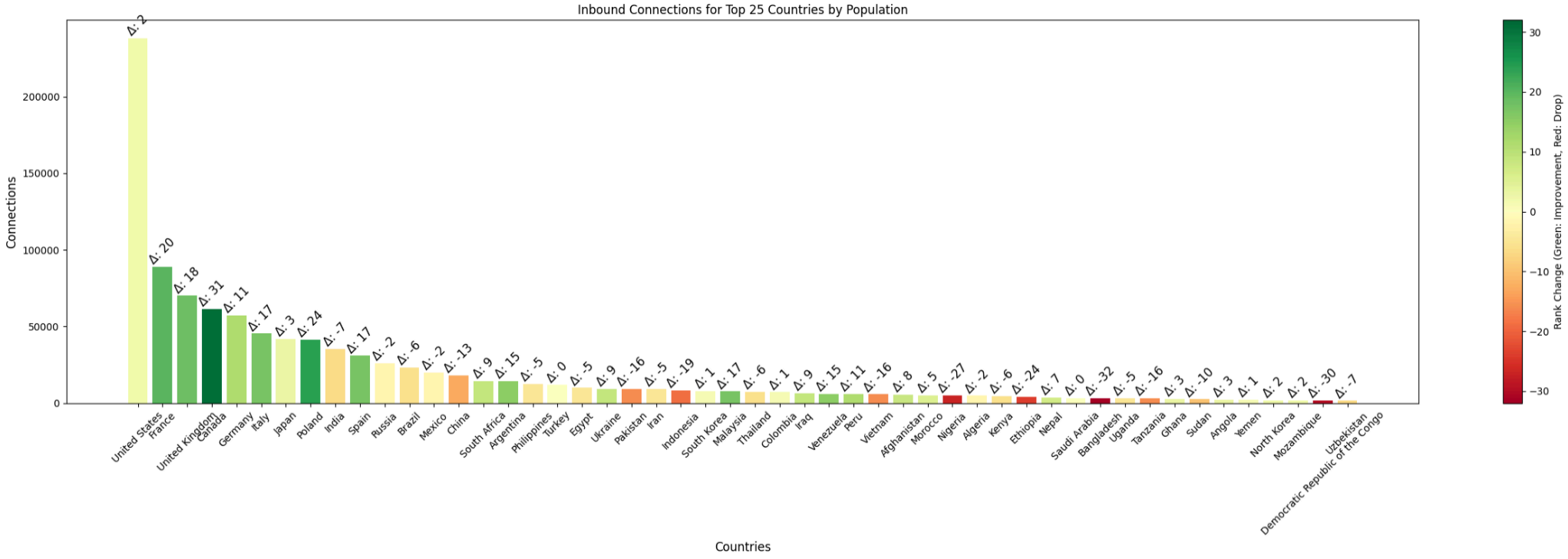

Fig 3. Difference from expected proportion of out degree by country

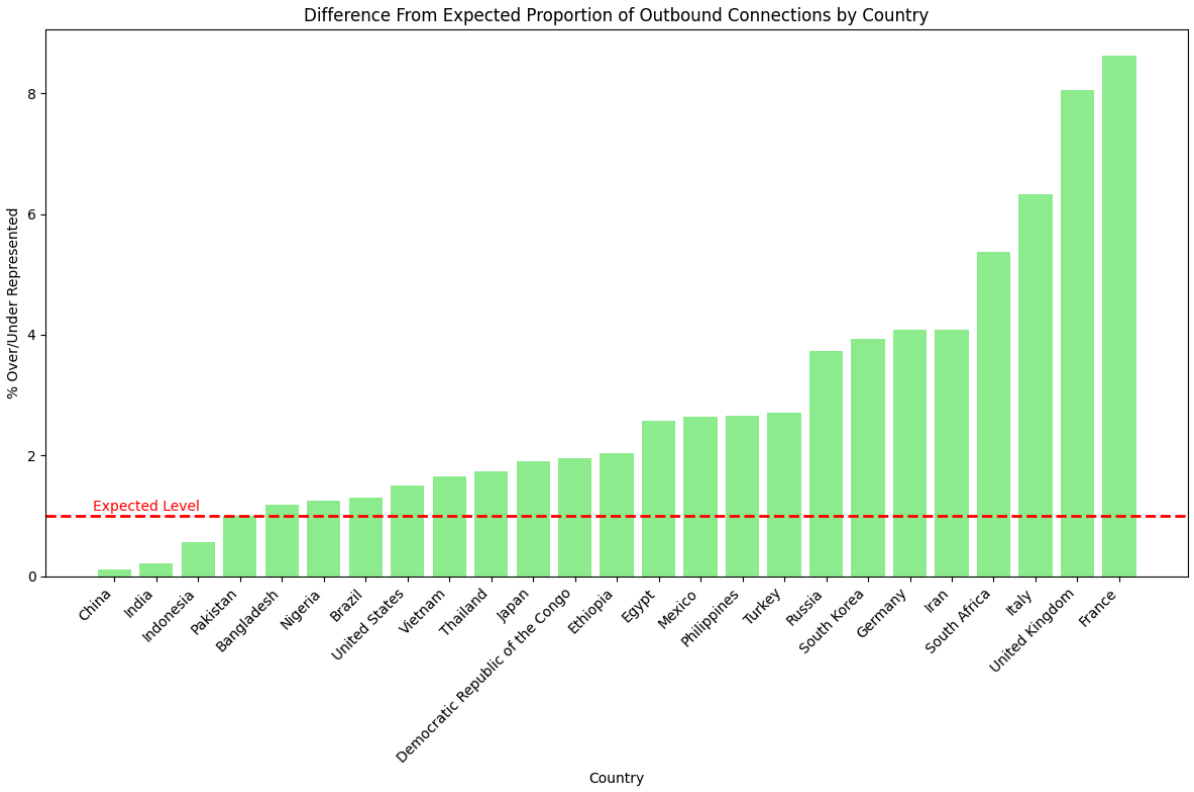

Fig 4. Difference from expected proportion of in degree by county