# Wikispeedia Network
# Graph Analysis and Link Prediction

Roidis Dimitrios

## Contents

# 1 Introduction

In this report, we explore the Wikispeedia dataset—a collection of user navigation paths through Wikipedia that has been cast into a directed graph, where each node represents an article and each edge a click-through link. Wikispeedia not only captures the topology of encyclopedic knowledge but also reflects how people actually traverse topics, making it an ideal testbed for studying information flow and human browsing intuition. By treating the network as directed, we preserve the asymmetry of hyperlinks—some pages funnel users onward, while others serve mainly as destinations.

After constructing and pruning the directed graph, we fix a two-dimensional layout for consistent visualization across analyses. We then apply a suite of directed-graph metrics—namely in- and out-degree distributions to gauge where attention concentrates, directed closeness to measure how efficiently articles can broadcast or receive traffic, and directed betweenness to identify key routing pages. Finally, by examining the underlying user behavior and efficiency, utilizing the paths used on the game.

# 2 Dataset and Preprocessing

The dataset comprises of tab-separated files: articles.tsv , categories.tsv, links.tsv containing directed hyperlinks between articles, and paths finished.tsv recording user navigation paths (though not directly used for visualization). Additionally, the shortest path distance is also included which pottentially turns useful to measure players' efficiency. A directed graph was constructed where nodes represent articles and edges represent hyperlinks, reflecting the underlying structure of Wikipedia's internal link network.

# 3 Structural Analysis and Related Metrics

To better understand the structural properties of Wikispeedia articles, we computed several metrics, mostly utilizing the networkx library: graph density, degree assortativity, in and out degree, pagerank, etc.

The structural properties of the Wikispeedia(Table 1) graph reveal key characteristics of its topology. The network comprises 4,592 nodes and 119,882 directed edges, indicating a densely interlinked collection of Wikipedia articles. Despite this density, the graph is fragmented into 519 strongly connected components, suggesting limited mutual reachability in parts of the graph,

| Metric | Value |
|---|---|
| Number of Nodes | 4592 |
| Number of Edges | 119882 |
| Strongly Connected Components | 519 |
| Weakly Connected Components | 2 |
| Graph Density | 0.0057 |
| Degree Assortativity | -0.0562 |
| Reciprocity | 0.2208 |

Table 1: Structural Metrics of the Wikispeedia Graph

whereas the presence of only 2 weakly connected components confirms a largely cohesive underlying structure. The graph exhibits low density (0.0057), consistent with typical real-world networks where most nodes are not directly connected. A negative degree assortativity coefficient (-0.0562) implies a tendency for high-degree nodes to connect with low-degree nodes, characteristic of hub-like behavior. Additionally, a reciprocity of 0.2208 reflects that approximately 22% of the edges are bidirectional, indicating moderate mutual linking between articles. These metrics collectively suggest a scale-free, navigable structure with dominant hub nodes and loosely coupled peripheral nodes.

| PageRank | | Betweenness | | In-Degree | | Out-Degree | |
|---|---|---|---|---|---|---|---|
| United_States | 0.0096 | United_States | 0.0941 | United_States | 1551 | United_States | 294 |
| France | 0.0064 | United_Kingdom | 0.0424 | United_Kingdom | 972 | Driving_on_the_left_or_right | 255 |
| Europe | 0.0063 | England | 0.0324 | France | 959 | List_of_countries | 244 |
| United_Kingdom | 0.0062 | Europe | 0.0270 | Europe | 933 | List_of_circulating_currencies | 236 |
| English_language | 0.0049 | Africa | 0.0242 | World_War_II | 751 | List_of_sovereign_states | 216 |

Table 2: Top 5 Centrality Measures for the Wikispeedia Graph

Regarding the rest of the metrics(Table 2), pageRank and betweenness centrality highlight different aspects of node importance. The United States consistently ranks highest in both PageRank (0.0096) and betweenness centrality (0.0941), indicating its dominance as both a well-connected and strategically central node in the network. France, Europe, and the United Kingdom also appear prominently, suggesting their frequent relevance in navigational paths, likely due to their geopolitical and historical significance. The inclusion of ”English language” and ”Africa” further reflects topics of broad interest and utility in user navigation. Mostly though, these results pinpoint to the true nature of wikipedia articles. Countries and concepts which act as an umbrella for a lot of other less significant concepts are more likely to be utilized in the search of an optimal path.

The in-degree and out-degree rankings provide additional insight into article connectivity. The United States again leads with the highest in-degree (1551) and a high out-degree (294), signifying that it is both widely referenced and contains numerous outbound links—supporting its role as a central hub. Articles such as ”Driving on the left or right” and ”List of countries” appear in the top out-degree rankings, indicative of list-based or reference-heavy content designed to interlink many topics. In contrast, historically and geographically significant entities like ”World War II”, ”Germany”, and ”London” dominate the in-degree rankings, reinforcing their role as endpoints or key references within the semantic structure of Wikipedia navigation. These metrics underscore how users traverse conceptual relationships in Wikispeedia, where centrality is shaped by both global relevance and link architecture.

## 4 Community Detection

Understanding the modular structure of Wikipedia is crucial for analyzing its semantics and the way users navigate in their search for information. We applied two widely used community detection algorithms: Louvain and Spectral Clustering. Both methods identify groupings of nodes (communities) that are more densely connected internally than externally.

| **Community** | **Size** | **Top Theme** |
|---|---|---|
| 1 | 843 | North_American_Geography |
| 2 | 823 | Countries |
| 3 | 789 | Science.Biology.Birds |
| 4 | 761 | Mathematics |
| 5 | 624 | Geography_of_Great_Britain |

Table 3: Sizes and Top Themes of Louvain Communities

The Louvain algorithm partitioned the Wikispeedia network into several densely connected “communities,” each of which corresponds to a coherent topical cluster of pages. As shown in Table 3 , the largest community (size 843) is centered on North American geography, reflecting tightly linked articles on U.S. states, Canadian provinces, and major physical features. Nearly as large, a second community (size 823) groups general “Countries,” capturing links between sovereign states, their capitals, and demographic or political topics. A slightly smaller third community (size 789) focuses on ornithology—articles about bird species, their habitats, and behaviors—highlighting how biological subfields form their own navigational clusters. Mathematics forms the core of the fourth community (size 761), bringing together entries on mathematical concepts, theorems, and biographical pages for notable mathematicians. Finally, the fifth community (size 624) is dedicated to the geography of Great Britain, tightly interlinking pages on the British Isles, counties, cities, and regional landmarks. Together, these five communities capture how readers naturally navigate between related topics, with each cluster reflecting both thematic depth and the density of inter-article hyperlinks.

| Community | Size | Top Theme |
|---|---|---|
| 1 | 2704 | Countries |
| 2 | 942 | Geography.Geography_of_Great_Britain |
| 3 | 450 | Science.Biology.Birds |
| 4 | 219 | Science.Chemistry.Chemical_elements |
| 5 | 184 | Science.Physics.Space_Astronomy |

Table 4: Sizes and Top Themes of Spectral Communities

Spectral clustering, on the other hand(Table 4), is a graph partitioning technique that leverages the eigenvectors of the Laplacian matrix of the network. By embedding nodes into a low-dimensional space defined by the smallest eigenvalues and then applying k-means clustering, this method captures global structural features that local methods might miss. As shown in Table 4, spectral clustering recovers several thematic communities—such as countries, British geography, and ornithology—that are consistent with those identified by Louvain. However, it also surfaces more specialized scientific topics, including clusters focused on chemical elements and astronomy, which Louvain did not isolate as distinct modules. This difference reflects spectral clustering's tendency to form tighter, globally separated partitions, even when they represent smaller portions of the graph. The largest spectral cluster (size 2,704) aggregates a broad range of country-related articles, while the remaining clusters are much smaller (ranging from 184 to 942 nodes), highlighting the algorithm's sensitivity to the global geometry of the graph. In contrast, Louvain produces more balanced partitions by dynamically identifying regions of high internal connectivity, resulting in mid-sized communities that correspond to broader but denser topical areas. It is worth noticing though, that different runs of Louvain, produced sightly different results. There was even an execution that produced exactly the same communities to spectral clustering. Louvain algorithm is non-deterministic by design, works iteratively and greedily and might converge on different local optima, especially for graphs with many near-optimal partitions.

On a more intuitive approach of spectral clustering, the first visualization (Figure 1) shows the Wikispeedia graph laid out with a spring force-directed algorithm and color-coded by the dominant theme of each spectral cluster. You can clearly see five visually coherent "blobs"—the large yellow cloud of Countries, which overlaps with the cloud of Geography of Great Britain the deep red island of Bird Biology and the orange cluster of Chemical Elements which overlaps with the cyan of Space Astronomy. Although the graph remains densely interconnected, the embedding + k-means pipeline teases out these global structural affinities, making thematic regions immediately legible by color. Interestingly, naturally though, communities of Chemistry and Physics, Countries and Great Britain overlap as they both refer to Countries and Physhics respectively. Even the birds cluster is overlaping with countries which makes perfect sense of the semantics of Wikipedia, as specific birds species are found in specific regions of the world.

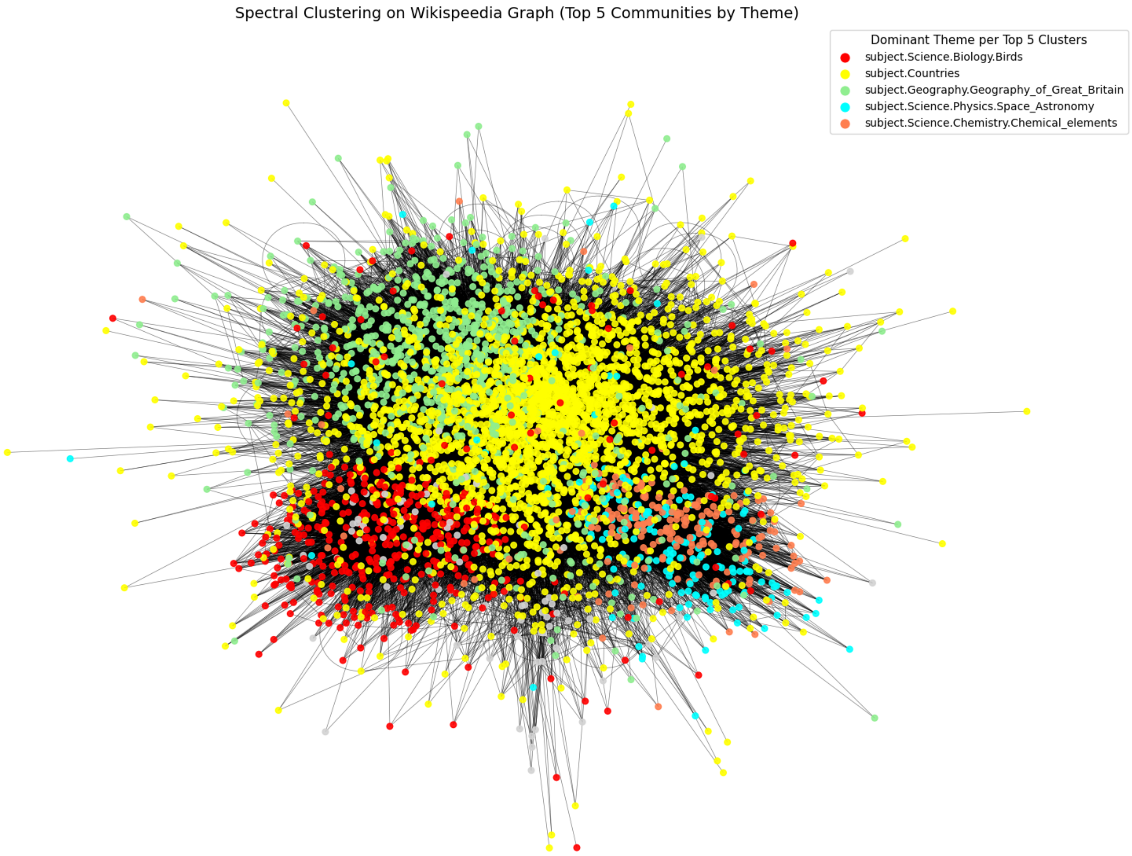

Figure 1: Wikispeedia Graph - Spectral Clustering Colored

The second visualization(Figure 2) uses a similar layout but instead sizes nodes by their PageRank score (we filtered to the top 4,000 articles and clipped outliers). Here, massive hubs like United States, World, and England dominate the center in bright magenta and purple, indicating their high navigational importance, while peripheral pages shrink to tiny cyan points. This view highlights who the “heavyweights” are in the encyclopedia’s link structure—pages that, in effect, “collect” the most random-walk traffic—rather than how pages group topically. Together, these two plots contrast community structure (spectral clustering colors) with node centrality (PageRank size), offering complementary perspectives on how readers traverse and cluster around different subjects in Wikispeedia.

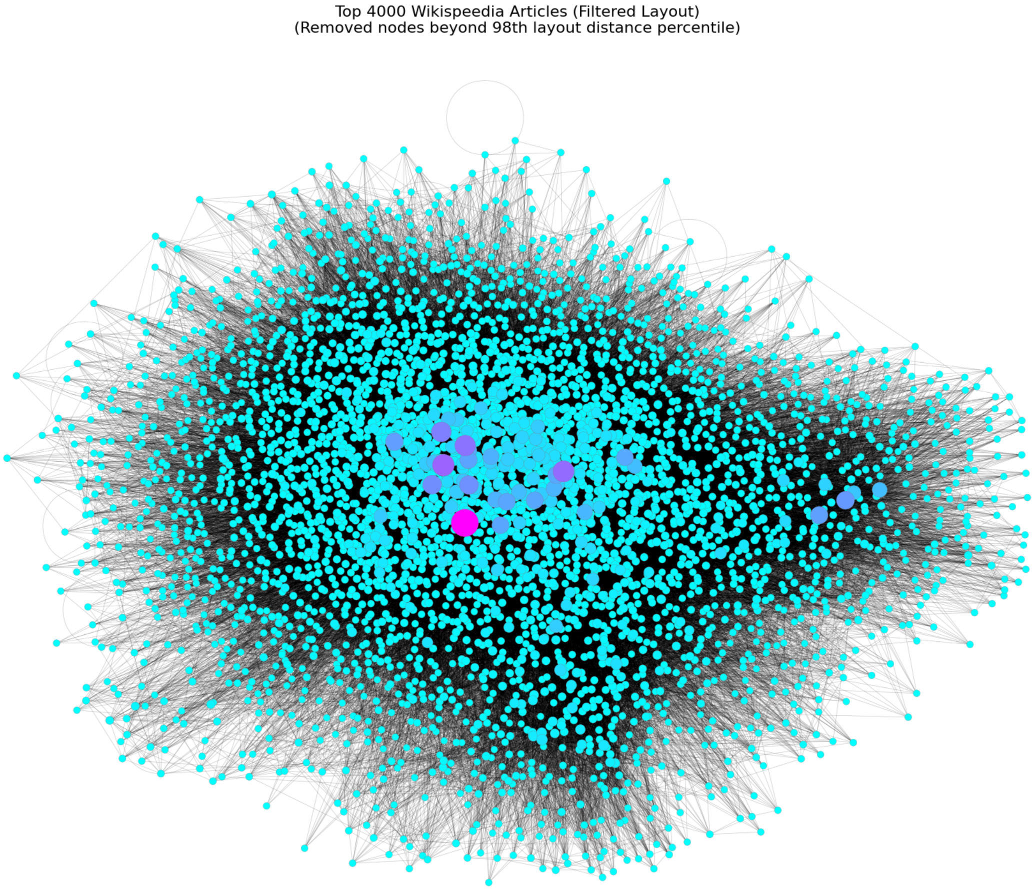
Top 4000 Wikispeedia Articles (Filtered Layout)
(Removed nodes beyond 98th layout distance percentile)

Figure 2: Wikispeedia Graph - Pagerank Filter

## 5 Players' Behavior and Efficiency

The path-based behavioral analysis reveals a rich tapestry of how players navigate Wikispeedia. Over 51,000 recorded navigation sessions averaged about 6.8 hops and roughly 2.5 minutes per path, suggesting that users typically pursue mid-length chains of associations before reaching their goal. At the outset, pages like Brain, Asteroid, and Theatre frequently serve as launch points—topics that are both concrete and evocative, beckoning players to spin off into related concepts. Conversely, destinations such as Telephone, Viking, and Zebra appear disproportionately often at the end of successful paths, indicating these entries act as common “sinks” or objectives in the game. Finally, the most visited pages overall—led United States, Europe, United Kingdom, and England dominate, reflecting the long-tail draw of geographic and cultural landmarks in player trajectories. Together, these statistics paint a picture of game dynamics where users begin with broad or intriguing topics, traverse semantic neighborhoods of intermediate concepts, and ultimately converge on well-connected or thematically resonant targets.

The efficiency metrics indicate that players’ actual routes are only about half as direct as the theoretical shortest paths—an average efficiency score(shortest path /actual path) of 0.528 means that, on average, users take roughly twice as many hops as the optimal route would require. Likewise, the mean detour(actual - shortest) of 3.39 extra hops shows that participants routinely meander through several off-topic or suboptimal nodes before homing in on their destination. Together, these figures suggest that while players often recognize broadly relevant connections, they frequently explore tangential links—perhaps in search of stronger semantic cues or simply out of curiosity—rather than following the most concise possible path. Finally the average duration of a game compared to its length is plotted(Figure 3) which reveals a mostly linear relationship between the number of hops a player takes and the total time they spend navigating: shorter paths (say 1–10 articles) tend to take only a few tens of seconds, while mid-length excursions (20–60 hops) average on the order of several hundred seconds. In other words, players spend roughly 20–25 seconds per click on average.

Beyond about 60–80 hops, however, the data become much noisier: there are only a handful of extremely long sessions (some exceeding 100 or even 400 hops), and those few outliers exhibit wildly varying durations—from under 200 seconds to nearly 3,000 seconds. This volatility at the tail likely reflects that very long games are both rare and driven by different behaviors (e.g. exploratory wandering, taking breaks, or multiple tab-hops).

Overall, the near-linear trend for the bulk of paths suggests that players maintain a consistent “per-hop” pace regardless of how long their journey becomes—reinforcing the picture from our efficiency analysis that actual detours add a roughly constant time penalty per extra hop.

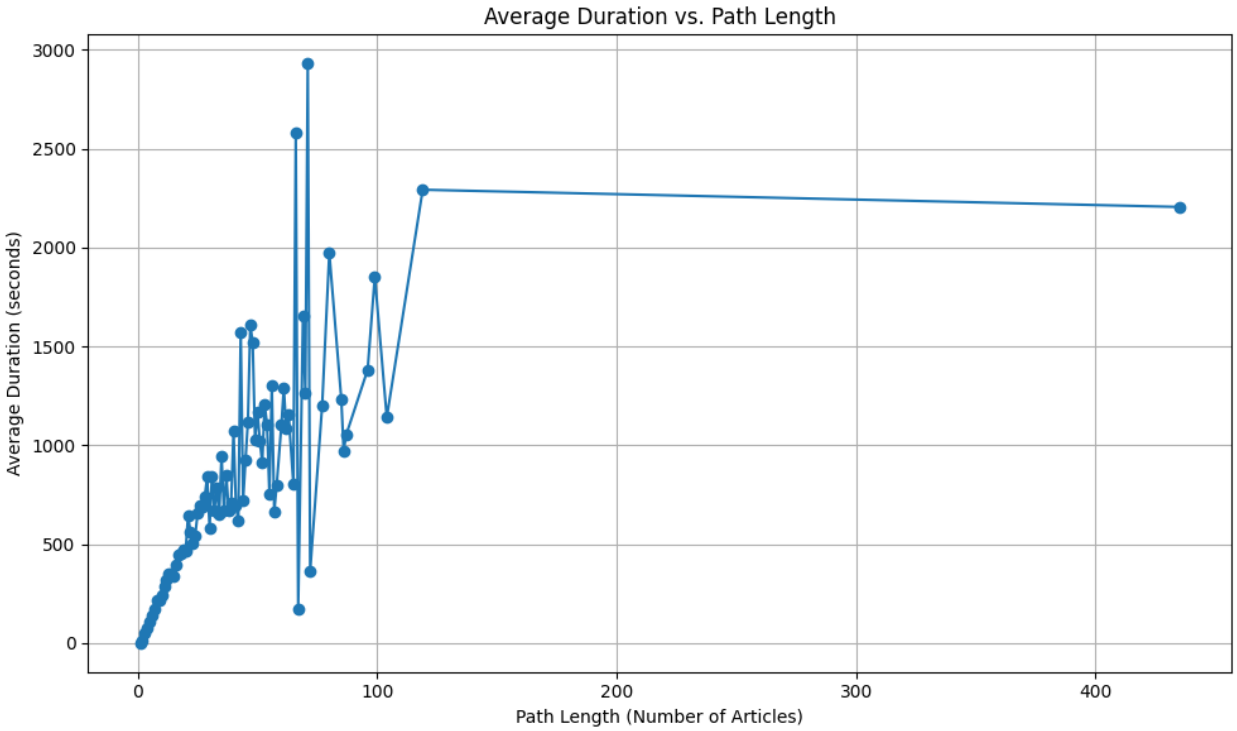

Figure 3: Average duration vs Path length

# 6 Link Prediction

Using the Jaccard coefficient on the training subgraph, we computed similarity scores for both true test edges and an equal number of randomly sampled non-edges. The resulting values range from 0 (no shared neighbors) up to approximately 0.16, reflecting the relative sparsity of the Wikispeedia hyperlink network. Notably, article pairs with strong topical or historical connections tend to receive the highest scores: for example, *Ancient_Egypt–Sumer* (0.136), *Brussels–Luxembourg* (0.139), and *History_of_South_Africa_in_the_apartheid_era–Desmond_Tutu* (0.125). These high-scoring pairs share many common neighbors—indicative of cohesive subtopics or closely linked historical contexts—so the Jaccard measure surfaces plausible missing links among semantically related pages.

However, even among these promising examples, the absolute scores remain relatively low, and some semantically weak or structurally idiosyncratic pairs also appear with nonzero values (e.g., *The_Beatles–Sony*, 0.041). This underscores that Jaccard purely quantifies neighborhood overlap without capturing thematic relevance. To improve predictive performance, we will compare these results against alternative heuristics (Adamic–Adar, Resource Allocation, and Preferential

Attachment) and evaluate precision–recall on the held-out test set. By combining multiple similarity metrics and tuning decision thresholds, we aim to better distinguish truly missing hyperlinks from random or spurious connections.

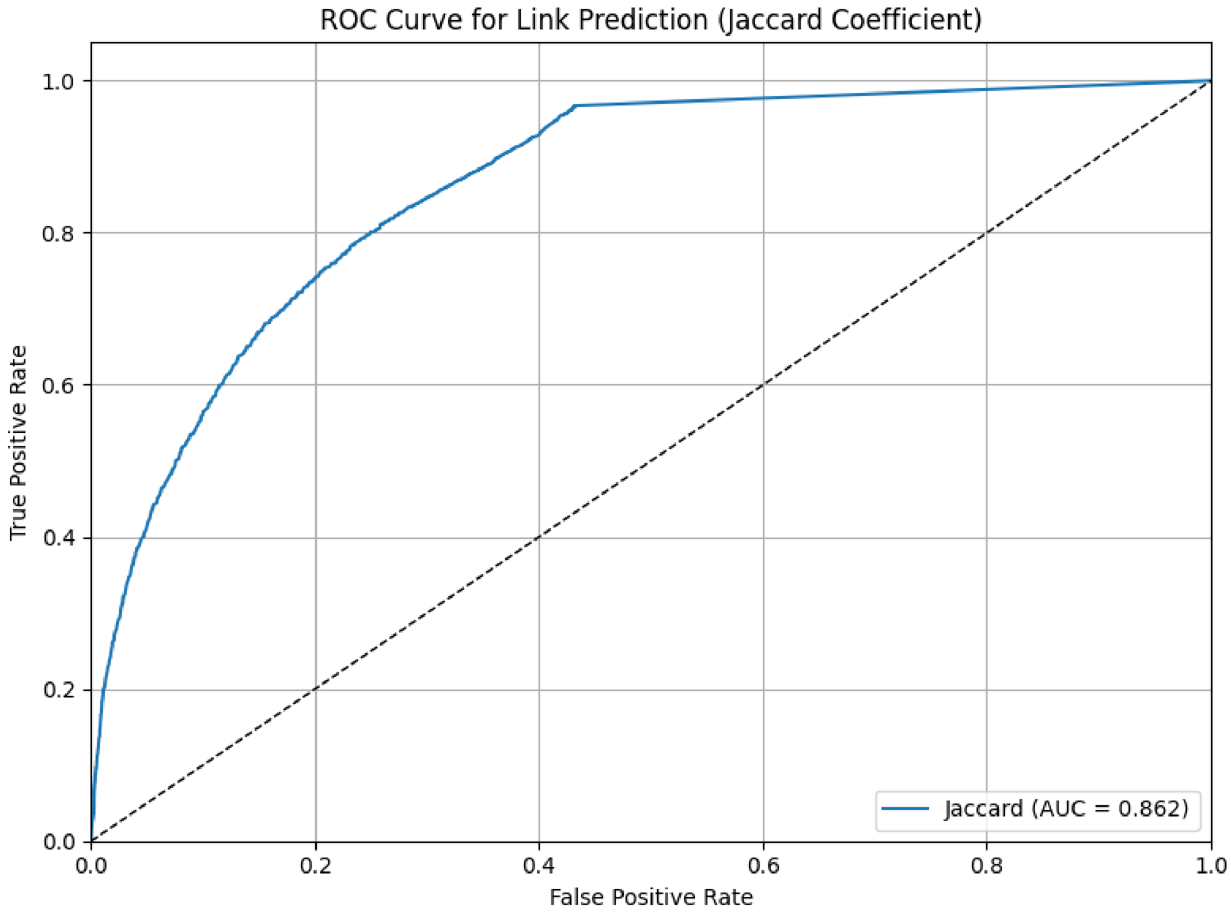

Figure 4: ROC Curve for Link Prediction with Jaccard Coefficient

In the context of Wikispeedia, the ROC curve for the Jaccard coefficient—yielding an AUC of 0.862—tells us that simple neighborhood overlap is a remarkably strong signal for recovering “missing” hyperlinks in the Wikipedia navigation graph. Concretely, if you pick one true hyperlink and one random non-link, Jaccard will rank the real link higher about 86% of the time. Moreover, the steep initial rise of the curve shows that you can capture a majority of actual links (e.g. ¿ 60% of them) while only introducing a small fraction of false positives (e.g. ¡ 10%). This means a Jaccard-based recommender could surface highly plausible link suggestions without overwhelming editors or readers with noise. At the same time, the flattening of the curve at high recall underscores diminishing returns—recovering the final few true links comes at the cost of many more spurious predictions—highlighting the value of combining Jaccard with complementary heuristics or semantic features to push beyond its baseline performance.

# 7 Conclusion

In this work, we constructed and analyzed a directed network representation of the Wikispeedia dataset to capture both the topology of Wikipedia and the actual browsing behavior of users. We began by measuring key structural properties (density, assortativity, reciprocity) and centrality metrics (PageRank, betweenness, in-/out-degree) to identify the most influential pages and assess overall navigability. We then applied two complementary community-detection methods—Louvain modularity and fixed-$k$ spectral clustering—to uncover major topical clusters in the network and compare how local versus global partitioning strategies shape community sizes and themes. Next, by mining over fifty-thousand real navigation paths, we quantified user efficiency, detours from shortest paths, and time per click, revealing consistent behavioral patterns and common start and end points. Finally, we evaluated link-prediction heuristics (Jaccard, Adamic–Adar, Resource Allocation, and Preferential Attachment) on a held-out test set, demonstrating that neighborhood-based measures can surface plausible missing hyperlinks—especially among closely related articles—while also highlighting their limitations in thematic relevance.

These analyses are important because they link the static structure of encyclopedic knowledge with the dynamic process of human exploration. Structural and community insights can guide the design of better navigation aids or link recommendations, behavioral metrics inform models of user search strategies and cognitive load, and link-prediction results point toward automated methods for suggesting new or missing connections. Together, they offer a unified framework for understanding how information is organized, how it is traversed, and how we might improve both the user experience and the underlying Wikipedia network.