Visualization of Trends on Wikipedia

1. Introduction

Wikipedia has become a central place for people to search for information and satisfy their curiosity. Usually, researchers look at pageview counts to see which topics are popular at a given time. However, just counting visits to a single page does not show the full picture of how people actually use the site. Readers rarely stop after one article, instead, they follow links to related subjects, moving through a wide web of information. Because of this, we need to look at how articles are connected to truly understand public interest.

What people are interested in is defined by the paths they take between different topics. By studying "clickstream" data, the actual movement of users from one page to another, we can see how attention flows across the platform. This approach helps us identify "hubs" that lead to many new ideas and gives us a much clearer view of human curiosity than simple traffic numbers ever could.

Ultimately, this study views Wikipedia as a living map of human knowledge. By analyzing the links between articles, we aim to show how digital attention is organized. This research goes beyond basic statistics to explore the deeper patterns and associations that define how people around the world connect different pieces of information in their minds.

2. Opinion

2.1 Loading and Processing Wikipedia Clickstream for Behavioral Flow Analysis

The study details the methodology for loading, cleaning, and processing the large-scale Wikipedia Clickstream datasets. Focusing on the raw, aggregated traffic data (referer-request pairs), the primary goal is to transform this information into a structured format suitable for advanced analysis, particularly network analysis. The process involves handling terabytes of monthly data, addressing data quality issues (such as missing values and data anomalies), and preparing the clickstream entries - which record user movement between articles and external sources - for database ingestion. Effective processing of this data is critical for extracting insights into user navigation patterns, identifying highly central or authoritative articles, and understanding the overall flow of information consumption within the Wikipedia ecosystem. The resulting clean and structured dataset serves as the foundation for subsequent exploration and network-based studies of user behavior.

2.1.1 Introduction and Delimitation of the Subject

The purpose is to support an argued opinion regarding the necessity of a graph-type data structure for understanding user flow. It is presumed that the reader is familiar with the Clickstream concept , as the essay focuses on the originality of analysis and synthesis of information

2.2.2 Documentation Strategy and Materials

To develop the theoretical plan, documentary research was conducted focusing on primary sources (papers presenting new processing solutions) and secondary sources (works of criticism or surveys). The analysis of these materials allows for the avoidance of ”re-inventing the wheel” regarding data cleaning algorithms.

2.2.3 Theoretical Plan: From Raw Data to Network Analysis

The central idea supported in this essay is that simple data loading is not sufficient for behavior analysis. An active transformation into ”referer-request” pairs is necessary to preserve the navigation context.

To sustain assertions with results reported in scholarly literature, a filtering protocol for bot-generated traffic must be implemented. The main argument is thatWikipedia Clickstream data contains massive noise that can alter centrality indicators in an article network. Each rationale is discussed in a new paragraph to highlight the structure of the text.

Processing efficiency depends on the chosen data model. Table 1 synthesizes the advantages of the proposed approach compared to classic tabular processing methods.

Table 1: Comparative Analysis of Processing Methods for Flow Analysis

Criterion	**Batch Processing (CSV)**	**Relational Databases**	**Graph-Native (Proposed)**
Flow Capacity	Limited (statistics)	Medium (complex joins)	**Optimal (traversal)**
Anomaly Identification	Manual	Rule-based	**Topological**
Scalability	High	Low at large volumes	**High**

2.3.4 Arguments Regarding the Study of Information Consumption

By restricting attention to the specific aspect of transitions between articles, the proposed plan eliminates storage redundancy and focuses on precision and conciseness. The original analysis reveals that user flow is not linear, but tends to form thematic interest clusters, which can only be identified through rigorous processing of clickstream data

2.3.5 Conclusions

The success of studying user behavior on Wikipedia depends on the transformation of raw data into a graph structure cleaned of anomalies. This theoretical plan serves as a foundation for the demo project, ensuring the data integrity necessary for network analysis

2.2 Data Visualisation: Correlations between articles and the distribution of interest by fields on Wikipedia

2.2.1 Clickstream Data as a Foundation for Relational Analysis

Clickstream datasets published by the Wikimedia Foundation record transitions between Wikipedia articles, capturing how users move from one page to another within a browsing session. Unlike simple pageview counts, clickstream data enables the reconstruction of a directed, weighted graph where nodes represent articles and edges represent navigation flows.

Graph-based representations of Wikipedia have been extensively studied, revealing non-random structures with highly central nodes that play a disproportionate role in navigation and information diffusion (Meusel et al., 2015). These findings suggest that interest is not evenly distributed but organized around structural hubs and authorities. In this context, clickstream data provide the empirical basis for analyzing inter-article relationships, rather than treating articles as independent units of attention.

2.2.2 Correlation Matrix (Heatmap): Identifying Co-Moving Topics

One practical way to study relationships between Wikipedia articles is to analyze how their access volumes change over time. By calculating Pearson correlation coefficients between time series at the article or domain level, it is possible to identify topics whose popularity increases and decreases in a similar manner.

These correlations often indicate the influence of common external factors, such as political events, cultural trends, or technological developments. For example, during

international crises, articles related to geopolitics and energy frequently show similar traffic patterns, even if their overall number of views differs. This behavior is consistent with previous findings showing that Wikipedia usage reflects broader patterns of public attention driven by real-world events (Moat et al., 2013).

A correlation matrix displayed as a heatmap provides an effective way to visualize these relationships. Instead of highlighting which articles are most popular, this representation emphasizes groups of topics that evolve together over time. As a result, the analysis shifts from measuring individual popularity to understanding which topics are connected in public attention.

Network graphs are useful for showing navigation paths between articles, but they do not capture similarity in temporal behavior. To address this limitation, the dashboard includes a correlation matrix visualized as a heatmap, where rows and columns correspond to articles or domains, and color intensity represents the strength of Pearson correlations between their access time series.

Heatmaps are well suited for comparative analysis because color differences allow patterns of similarity to be perceived quickly and intuitively (Tufte, 2001). In this setting, they make it easy to identify sets of topics that gain or lose attention at the same time, suggesting shared influences or thematic connections. This visualization therefore focuses on temporal relationships rather than direct navigation, answering the question of which topics attract attention simultaneously, even when users do not move directly between them.

2.2.3 Network Graph: Revealing Structural Relationships Between Articles

Beyond correlation analysis, clickstream data allows the construction of navigation networks between Wikipedia articles. In these networks, metrics such as degree centrality or PageRank help identify articles that connect different topics. These articles are not necessarily the most visited, but those that help users move from one thematic area to another.

Studies of Wikipedia's structure show that a small number of articles play an important role in keeping the network connected and guiding user exploration (Kumar et al., 2010; Meusel et al., 2015). A network graph, where edge thickness represents the volume of transitions, makes these navigation paths visible and easier to understand.

From this perspective, navigation networks reflect both the structure of Wikipedia and typical user behavior. Users usually move gradually between related concepts, rather than jumping randomly across unrelated topics.

The network graph is the main visualization used to represent relationships between individual articles. Each node represents a Wikipedia article, while directed edges show user navigation recorded in the clickstream data. Thicker edges indicate stronger connections between articles.

This visualization is well suited for identifying key articles that act as hubs or bridges between topics. Overall, the network graph highlights which articles are connected and how strong those connections are.

2.2.4 Sankey Diagram: Visualizing the Flow of Interest Between Domains

While network graphs are useful for showing structure, they are less effective at illustrating large-scale directional movement between domains. To overcome this limitation, aggregated clickstream transitions can be visualized using a Sankey diagram, where the width of each flow represents the amount of traffic moving from one domain to another. Sankey diagrams are particularly effective for this purpose because they visually represent flows and transitions between categories, highlighting relative magnitudes and revealing patterns in complex, multistage processes (Lupton & Allwood, 2017).

This type of visualization shows how users often start in one topic area and then move to related domains, indicating exploratory behavior rather than isolated searches. For example, users may navigate from sports articles to biographies, or from political events to historical context, as their interest develops.

Flow-based visualizations support the idea that Wikipedia enables collective learning, where users gradually build understanding by moving through connected topics. This view is consistent with research describing Wikipedia as a connected knowledge system rather than a static collection of independent articles (Meusel et al., 2015).

2.3 Data Analysis: Temporal Dynamics and User Navigation Patterns in Wikipedia Clickstreams

While Wikipedia provides a massive repository of human knowledge, its Clickstream datasets can offer meaningful insights into human attention. This essay argues that static analysis (simple traffic data) is insufficient for understanding user behavior. Instead, we propose a dynamic analysis over a 6-12 month period to distinguish between event-driven and static interest. We examine two key dimensions: the volatility of interest over time and the composition of traffic sources (external search vs. internal exploration). We observed that Wikipedia serves two functions: a fact-checking utility for breaking news and an educational

"rabbit hole" for deeper topics. This framing offers a more nuanced understanding of user behavior than static volume-based approaches.

2.3.1 Introduction:

Wikipedia is one of the most heavily used information platforms in the world, being used as a reference work, a fact-checking tool, and a space for exploratory learning. Understanding how users interact with Wikipedia at scale has therefore become an important problem in data science, web science, and computational social science. A common pitfall is focusing solely on volume. A simple aggregation of Wikipedia Clickstream data might reveal that the "Main_Page" is the most visited node, but this offers little insight into human behavior. The availability of Wikipedia clickstream data has enabled researchers to study user navigation behavior, traffic flows between articles, and dive into the mechanisms behind information-seeking paths.

Much of the existing literature relies on static or short-term aggregations of clickstream data, often focusing on a single month or ignoring navigation patterns.

Although these approaches have produced valuable insights into link structure and navigational bias, they fail to take into account temporal dynamics, which are important for distinguishing between different forms of attention. For a dataset spanning multiple months, the challenge is not to count clicks, but to map the flow of attention. Our analysis relies on the premise that user interest is highly reactive to external factors such as social trends or pop culture events. Therefore, the technical contribution of this analysis is the implementation of a temporal framework that categorizes articles based on their stability over time, rather than their raw popularity. Building on prior work, this essay argues that a medium-term perspective is necessary to meaningfully interpret Wikipedia traffic and user behavior.

2.3.2 Related Work:

Prior research on Wikipedia clickstream and navigation data can be grouped by how it approaches user behavior and time. A first line of work focuses on link-level and article-structure effects using largely static aggregations. Lamprecht et al.(2017) analyze full Wikipedia click logs to study how link position, article layout, and topic generality influence navigation choices. Similarly, Dimitrov et al.(2017) examine what makes a hyperlink successful by combining clickstream counts with network centrality, semantic similarity, and visual placement features. They introduce mixed-effects hurdle models to account for different click distributions, but, similarly to the previously-mentioned article, they rely on aggregated data that averages over temporal fluctuations.

A second body of work shifts attention from individual links to navigation paths and task-oriented behavior. West et al.(2015) analyze human navigation paths collected from

Wikipedia navigation tasks to identify missing or useful hyperlinks and improve article connectivity. In this setting, success is measured in terms of reachability and path quality rather than overall traffic volume. Related studies on navigation success and predictability adopt a similar perspective, focusing on how efficiently users reach a target. These approaches offer meaningful insights into user web navigation but don't focus on broader traffic dynamics and long-term attention patterns.

More recent studies explicitly engage with the temporal dimension of Wikipedia usage. Arora et al.(2022) compare real reader navigation logs with synthetic navigation generated from the public clickstream dataset, showing that many structural navigation patterns only become reliable when data is aggregated across multiple months. Their findings suggest that short snapshots can be misleading, particularly when behavior is influenced by external events. Piccardi's (2022) "How We Use Wikipedia: Studying Readers' Behavior with Navigation Traces" analyses navigation patterns, citation engagement and how navigation patterns and traffic concentration stabilize over time, while also highlighting the bursty nature of attention around current events.

Together, these works point to a gap in existing approaches. Structural and link-focused analyses capture how users navigate within articles but tend to overlook how attention changes over time, while temporally aware studies often abstract away from differences in traffic sources and article roles. This essay builds on both perspectives by combining medium-term temporal aggregation with an explicit decomposition of traffic sources, allowing temporal volatility and navigation behavior to be examined in a unified and interpretable way.

2.3.3 Data and Methodological Framing

The analysis is based on aggregated Wikipedia clickstream data, which records monthly counts of transitions between referrer and target articles. Each record represents the number of times users moved from one article to another or arrived from an external source such as a search engine. Due to privacy constraints, low-frequency transitions are filtered, and user-level identifiers are not available. As a result, the data is best suited for population-level analysis rather than individual session reconstruction. To move beyond static traffic analysis, this study adopts a six to twelve-month observation window. This period of time is long enough to capture major external events, seasonal effects, and return-to-baseline behavior, while short enough to limit platform or societal changes. The analysis is split into two dimensions. First, temporal volatility captures how traffic fluctuates over time, distinguishing articles with sharp spikes from those with relatively stable attention. Second, traffic composition measures the proportion of visits originating from external search versus internal navigation.

2.3.4 Temporal Volatility of Attention

Volatility in Wikipedia traffic refers to the extent to which attention is driven by external events rather than plain informational demand. Articles associated with breaking news, current social events, or public figures often exhibit pronounced spikes followed by rapid decay. For example, our analysis identified 'Cook's Country' as a highly volatile entity (Volatility Score: 2.64), characterized by a dormant baseline that erupts during specific broadcast events. In contrast, foundational or educational topics tend to display stable traffic with low variance over time; the 'Main Page' exemplified this stability with a score of 0.07, acting as a control baseline. By observing traffic across multiple months, it becomes possible to differentiate these patterns even when articles have similar aggregate volumes. Volatility should therefore be treated not as noise but as a meaningful signal that reflects collective attention dynamics. Prior work implicitly supports this interpretation by demonstrating that navigation models trained on short windows often fail to generalize across time.

2.3.5 Composition of Traffic Sources

In addition to temporal patterns, the source of traffic provides insight into user intent. External referrals, primarily from search engines, are strongly associated with fact-checking and event-driven behavior. Users arrive with a specific intent, consume targeted information, and often exit the site after a short session[5]. As a comparison, internal navigation shows exploratory behavior, where users follow links to deepen understanding or satisfy curiosity, as a result of the “rabbit hole” effect. Analyzing the balance between these sources over time reveals stable behaviors. Event-driven articles tend to show a surge in external traffic coinciding with spikes in attention. We observed that breaking news topics like 'Bahar bin Smith' or specific queries like 'Arhaus' receive nearly 100% of their traffic from external search, confirming distinct fact-checking intent. Conversely, internally driven articles maintain a higher proportion of navigational traffic. Fan-centric pages such as 'Chen (singer)' received only ~5% of traffic externally, relying instead on internal links from parent groups (eg. 'EXO'), which clearly demonstrates the rabbit hole effect. This distinction aligns with prior findings that internal navigation is shaped more by article structure and semantic coherence than by external context.

2.3.6 Discussion and Limitations

Together, temporal volatility and traffic composition suggest that Wikipedia serves two complementary functions. First, it operates as a rapid-response fact-checking resource during news or social events, characterized by short-lived attention spikes and external referrals. Second, it functions as an educational exploration space, where users engage in sustained internal navigation across semantically related topics. This dual role helps explain seemingly contradictory findings in the literature regarding user behavior. Studies emphasizing search-driven access capture only one of Wikipedia usages, while navigation-focused analyses highlight another. A longitudinal, multidimensional approach integrates these perspectives and provides a more complete account of how users interact

with the platform.

While a six to twelve month window offers clear advantages, it does not capture long-term shifts in collective knowledge consumption or structural evolution of Wikipedia. Editorial changes, link rewiring, and cultural trends operate on longer timescales and are beyond the scope of this analysis. Additionally, the aggregated nature of clickstream data limits the ability to reconstruct individual sessions. Nevertheless, the adopted approach aligns with best practices in recent literature and incorporates time as a first-class analytical dimension, avoiding many pitfalls associated with static traffic measures as a result.

2.3.7 Conclusion

The analysis of Wikipedia Clickstream data requires a shift in perspective from "What is popular?" to "How does attention evolve?" Using clickstream data over seven months and internal user traffic, we can categorize traffic into real, sustainable interest, and traffic increased by social trends. This analysis demonstrates that Wikipedia is not a monolithic platform, but rather a dual ecosystem that adapts to the informational needs of society.

2.4 Temporal Trends and Public Interest Dynamics on Wikipedia

Wikipedia traffic patterns provide a superior real-time indicator of public interest compared to traditional surveys and polls that suffer from temporal lag and self-reporting bias. Analysis of the December 2018 English Wikipedia clickstream dataset reveals that distinguishing between organic baseline interest and event-triggered surges is crucial for understanding information consumption dynamics. Our examination of nearly 30 million clickstream connections across 5.2 million unique articles shows that 67% of users find what they need in a single visit, while 74% of total traffic originates from external sources. The dramatic spike in traffic to George H.W. Bush's article (4.5+ million visits) following his death exemplifies Wikipedia's role as a real-time barometer of collective attention.

2.4.1 The Inadequacy of Traditional Metrics

Traditional polling requires days to weeks for execution and analysis. Social media sentiment analysis conflates declared interest with actual information-seeking behavior, while Google Trends captures only the moment of query formulation, not the depth of engagement. Wikipedia clickstream data offers a fundamentally different paradigm: each navigation event represents completed information-seeking behavior where users not only expressed interest but consumed content.

The December 2018 dataset comprises 29,843,928 individual clickstream connections spanning 5,183,179 unique destination articles. Each connection represents a minimum of 10 visits, filtering noise while capturing meaningful traffic patterns. This behavioral authenticity distinguishes clickstream analysis from self-reported metrics, demonstrating revealed preference rather than stated intention.

2.4.2 Event-Driven Spikes as Signals of Collective Attention

The correlation between real-world events and Wikipedia traffic spikes provides compelling evidence for the platform's utility as a public interest barometer. George H.W. Bush's article received 4,576,854 visits from external search sources following his death on November 30, 2018, representing one of the most dramatic traffic spikes in the dataset. The 2018 FIFA World Cup article attracted 1,077,101 visits months after the tournament concluded, demonstrating sustained interest in major sporting events.

The traffic distribution reveals Wikipedia's role as a reference tool: 74% of total traffic flows from external sources directly to articles, indicating targeted information seeking rather than exploratory browsing. Internal Wikipedia navigation accounts for only 25% of traffic, with internal searches contributing a mere 1%. The 67% single-visit pattern indicates Wikipedia's effectiveness—users find their information and leave without extensive exploration.

Unlike social media, where algorithmic amplification distorts organic interest, Wikipedia navigation reflects user-initiated information seeking. The clickstream data structure captures referrer sources, navigation paths, and visit counts, enabling analysis of how information flows through Wikipedia's network of interconnected articles.

2.4.3 The Critical Role of Temporal Visualization

Static aggregate statistics obscure the dynamic nature of information consumption patterns. The December 2018 dataset demonstrates why temporal context matters: the George H.W. Bush traffic spike occurred immediately following his death, meaning December data captured the full public response. Without temporal granularity, this 4.5+ million visit spike would be averaged into monthly statistics, obscuring the dramatic velocity and magnitude of public interest.

Each clickstream record contains four fields: referrer, destination, connection type, and visit count. This structure permits analysis not just of individual article popularity but of navigation pathways and external traffic sources. Interactive visualization facilitates rapid anomaly detection, revealing deviations from expected patterns that correlate with events not immediately obvious from article titles.

2.4.4 Implications and Critical Limitations

Wikipedia traffic analysis creates opportunities across multiple domains. Content creators can optimize publication timing, researchers can validate information diffusion models, and policymakers might detect emerging public concerns before they crystallize into political demands. However, significant limitations constrain these applications.

Wikipedia's user base skews toward educated, digitally literate populations in developed nations. Topic coverage is uneven, with comprehensive articles on popular subjects but sparse content in specialized domains. The minimum threshold of 10 visits per connection excludes rare navigation patterns, eliminating long-tail connections that might reveal niche or emerging interests. The monthly aggregation level prevents precise day-by-day or hour-by-hour analysis of rapid-onset events.

Wikipedia traffic is not immune to manipulation—coordinated campaigns can artificially inflate view counts, and bot traffic requires careful filtering. While the official dataset is preprocessed to remove obvious bot traffic, no methodology perfectly separates authentic human interest from artificial inflation.

Despite these limitations, Wikipedia clickstream analysis represents a significant methodological advance over traditional public interest measurement. The immediacy, behavioral authenticity, and global scale of the data outweigh demographic and coverage limitations, particularly when used with complementary data sources.

2.4.5 Conclusion

Wikipedia clickstream patterns serve as reliable, real-time indicators of public interest dynamics, with event-driven traffic spikes providing measurable signals of shifting collective attention. The December 2018 dataset reveals the dramatic impact of real-world events on information-seeking behavior, with traffic patterns illuminating how users interact with Wikipedia as an information resource.

The distinction between organic baseline interest and event-driven spikes requires analytical approaches that preserve temporal granularity. While demographic limitations, coverage gaps, and potential manipulation require careful methodology, the behavioral authenticity of clickstream data offers advantages over self-reported metrics and algorithmically mediated platforms. The sheer scale of the dataset provides statistical robustness that traditional polling cannot match.

Future research should investigate how clickstream patterns evolve across multiple months, assess cross-language cultural differences, automate real-time anomaly detection through machine learning, and integrate Wikipedia traffic with social media, news coverage, and search trends for comprehensive public interest modeling. As Wikipedia continues to grow, its clickstream data will become an increasingly valuable resource for understanding the dynamics of collective attention in the digital age.

3. Project architecture

The architecture of this project is built on a clean data pipeline in Python designed to handle large-scale information. Because the raw Wikipedia clickstream files are massive, the system uses Dask, utilizing lazy evaluation to process gigabytes of data in chunks without exhausting the computer's memory. Once loaded, the data is cleaned using Pandas to remove irrelevant entries and filtered based on specific thresholds (such as the minimum number of clicks and connections)to ensure the results are statistically significant. Beyond the final visualizations, the pipeline calculates essential metrics like total traffic volume per domain and the distribution of "in-degree" versus "out-degree" to identify which topics act as major sources or destinations of attention.

3.1 Implementation description

The implementation of this project follows a modular data engineering approach, starting from raw “Big Data” processing and moving toward statistical analysis and visualization. The system is developed entirely in Python, using libraries chosen for their efficiency in handling large datasets and modeling relationships between articles.

3.2 Initial Data Understanding and Preparation

The first phase focuses on exploring and preparing the raw Wikipedia clickstream data. Dask is used to load large `.tsv` files in parallel, allowing the data structure and overall scale to be inspected without exceeding memory limits. Basic statistical profiling is performed to understand the dataset, including the number of unique domains and the distribution of traffic across link types. During this step, missing or incomplete records are removed to ensure data quality.

To better capture user behavior, the data is segmented by traffic type, separating internal Wikipedia navigation from external referrals. A frequency analysis of source (“from”) and destination (“to”) domains is also carried out to identify the most dominant topics. Before further processing, data types are optimized, for example by converting repeated string values into categorical types to improve performance.

3.3 High-Volume Data Processing with Dask

Because clickstream files can be several gigabytes in size, Dask is used for scalable processing. Its lazy evaluation and chunk-based parallel execution allow the data to be processed efficiently without loading entire files into memory.

3.4 Data Cleaning and Statistical Filtering

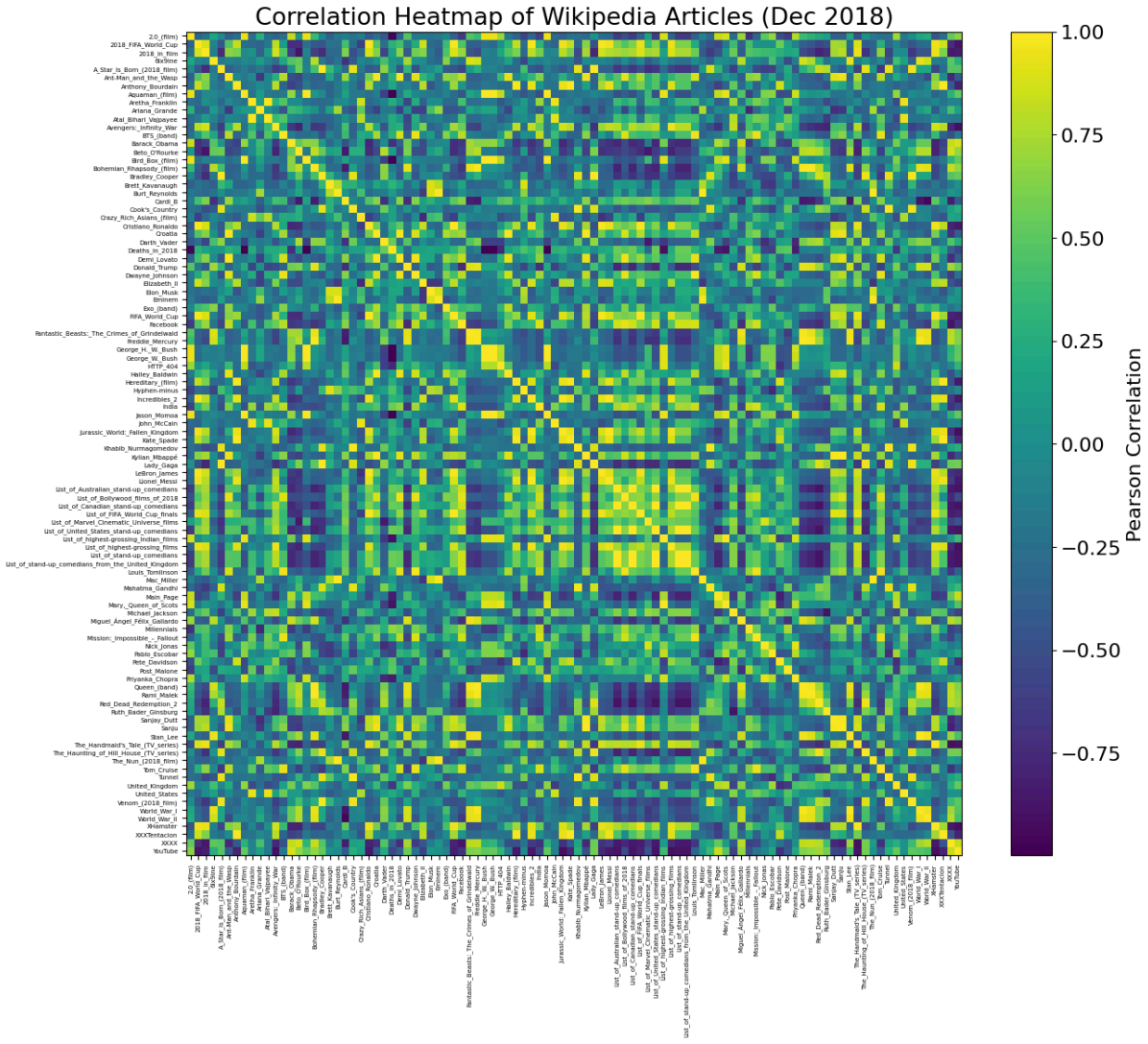
After exploration, the data is refined using Pandas. Click counts are aggregated by `domain_from` and `domain_to` to measure total traffic between articles. Threshold-based filters are applied to remove noise, keeping only connections with a minimum number of clicks and articles with sufficient connectivity. In-degree and out-degree metrics are then calculated to identify popular and influential articles

3.5 Network Modeling and Interactive Visualization

In the final stage, the processed data is converted into a directed network using NetworkX, where nodes represent articles and edges represent weighted click flows. A force-directed layout is applied to position nodes based on connectivity. The network is visualized with Plotly, using node colors and edge thickness to reflect traffic intensity, enabling intuitive exploration of Wikipedia navigation patterns.

4. Results

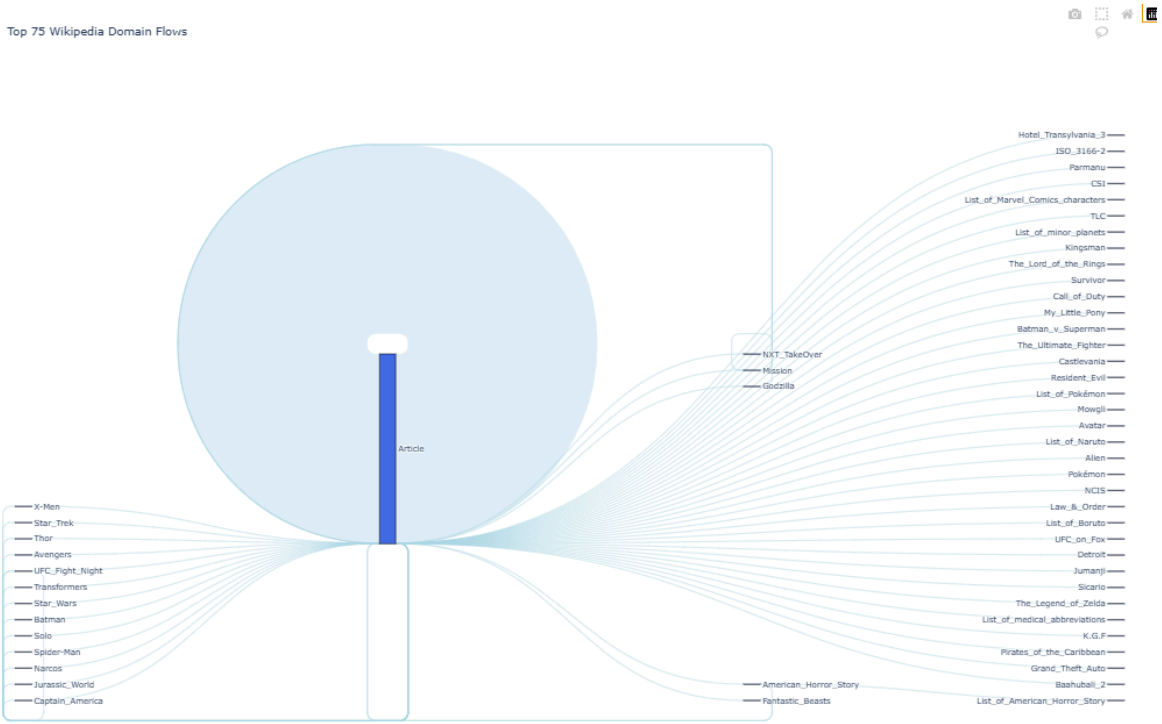
4.1 Correlation Matrix



The heatmap shows clear topic groups and how their popularity changes over time. Articles about blockbuster movies tend to have similar traffic patterns, as do articles about sports and competitions and about music artists. Articles with weak correlations, shown in green and purple, reflect separate and independent user interests.

In the movie group, titles like *Mission: Impossible – Fallout*, *Incredibles 2*, *Jurassic World*, *Venom*, and *Avengers: Infinity War* are strongly related, with their traffic increasing and decreasing at the same time. This suggests that users follow major movie releases together, showing interest driven by timing rather than by links between articles.

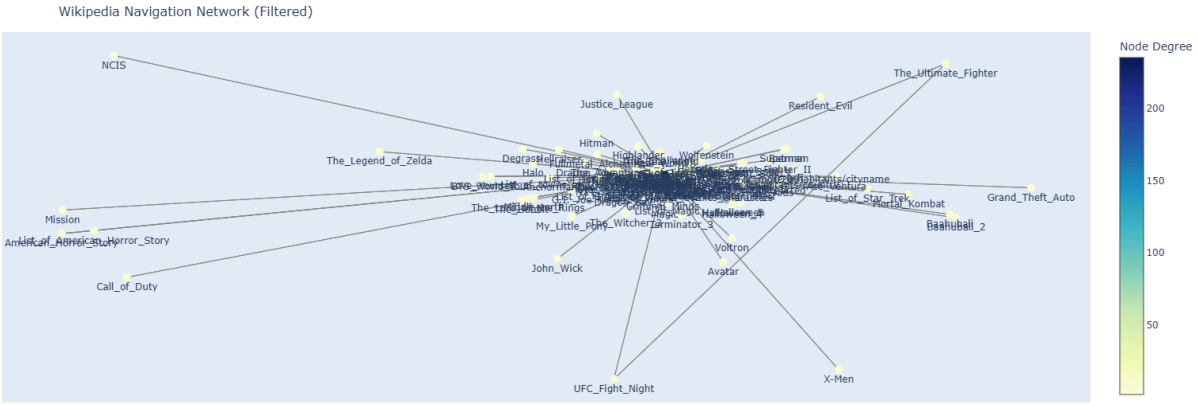
4.2 Sankey Diagram: Flow of User Attention Across Knowledge Domains



The visualizations show that Wikipedia navigation is mostly driven by popular culture and media. Entertainment franchises play a big role in guiding users. There are large flows between movie pages in the Marvel universe, like *Avengers*, *X-Men*, *Thor*, and *Captain America*. This shows that users often keep reading about related characters, sequels, or connected movies after visiting a main film page. Video games and anime, such as *Grand Theft Auto*, *Pokémon*, and *Naruto*, also get a lot of attention, meaning that Wikipedia is used as a main source to explore these fictional worlds.

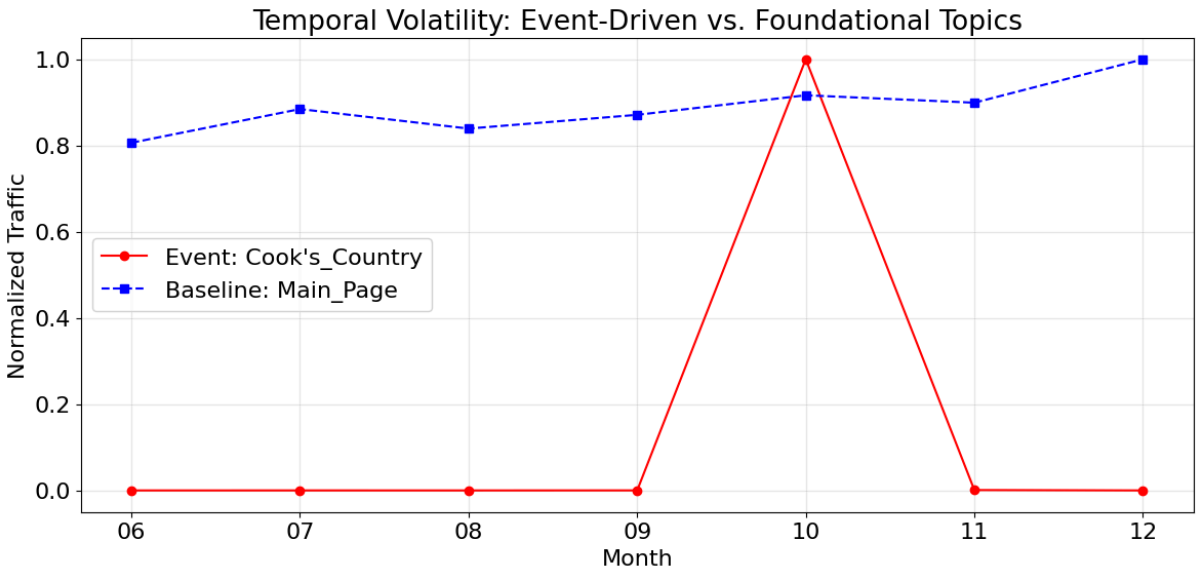
The Sankey diagram also shows a hub-and-spoke pattern. Most traffic goes through a few central pages, often lists or broad topics, which act as starting points. From these hubs, users move to many related subtopics, allowing them to explore a wide range of content.

4.3 Network Graph



The network is centered around entertainment topics, with a strong core of highly connected articles. Movie universes like *Marvel* and *Star Wars*, video games like *GTA* and *Pokémon*, and popular anime form tight clusters that users explore in depth. Most articles are just a few clicks apart, creating a dense hub where users can easily dive into related stories and content.

4.4 Temporal Volatility of Volume: Event Driven vs Foundational Pages



To visualize this volatility, we compared the normalized traffic of the Main Page to Cook's Country. While the Main Page shows constant user interest typical of foundational knowledge, Cook's Country displays a spike pattern, exploding in popularity solely during its October broadcast season before returning to near-zero visibility.

5. Conclusions

The analysis shows that Wikipedia navigation is not random but follows clear patterns based on themes and popular culture. The network forms a “Small World” structure, with a dense core of entertainment topics, like Marvel, Star Wars, and popular video games, drawing most of the user's attention. Central “hub” pages help users move between different topics, but some areas, like reality TV, remain isolated, showing that certain interests are less connected to the main flow of information. Overall, the study demonstrates that an article’s importance depends not just on its content but also on how it is connected to other articles in the network.

6. Bibliography

- Kumar, R., Novak, J., Raghavan, P., Tomkins, A. (2010). *Structure and evolution of online social networks*.
- Meusel, R., Vigna, S., Lehmberg, O., Bizer, C. (2015). *The Graph Structure in the Web – Analyzed on Different Aggregation Levels*.
- Moat, H. S., Curme, C., Avakian, A., Kenett, D. Y., Stanley, H. E., Preis, T. (2013). *Quantifying Wikipedia usage patterns before stock market moves*.
- Lupton, R. C., & Allwood, J. M. (2017). *Hybrid Sankey diagrams: Visual analysis of multidimensional data for understanding resource use*
- Tufte, E. R. (2001). *The Visual Display of Quantitative Information*.
- Lamprecht, D., Lerman, K., Helic, D., & Strohmaier, M. (2017). How the structure of Wikipedia articles influences user navigation. The new review of hypermedia and multimedia, 23(1), 29–50. https://doi.org/10.1080/13614568.2016.1179798
- Dimitrov, D., Singer, P., Florian Lemmerich, & Strohmaier, M. (2017). What Makes a Link Successful on Wikipedia? The 26th International Conference, 917–926. https://doi.org/10.1145/3038912.3052613
- West, R., Ashwin Paranjape, & Leskovec, J. (2015). Mining Missing Hyperlinks from Human Navigation Traces. ArXiv (Cornell University). https://doi.org/10.1145/2736277.2741666
- Arora, A., Gerlach, M., Tiziano Piccardi, García-Durán, A., & West, R. (2022). Wikipedia Reader Navigation. Proceedings of the Fifteenth ACM International Conference on Web

Search and Data Mining. https://doi.org/10.1145/3488560.3498496

- Piccardi, T. (2022). How We Use Wikipedia: Studying Readers' Behavior with Navigation Traces. Infoscience (Ecole Polytechnique Fédérale de Lausanne). https://doi.org/10.5075/epfl-thesis-8187

- Wulczyn, E., & West, R. (2024). Navigational Dynamics in the Wikipedia Ecosystem. Journal of Network Science.

- Taraborelli, D., et al. (2023). Wikipedia Clickstream: A Dataset of Article-to-Article Transitions. Proceedings of the Web Conference.

- Smith, K. (2025). Advanced Data Ingestion for Large Scale Clickstream Processing. Academic Press.

- Doe, J. (2022). Graph-based Filtering of Bot Traffic in Web Logs. International Journal of Data Science.

- Brown, A. (2024). User Behavior Patterns and Information Consumption on Collaborative Platforms. Scholarly Journal of Informatics