An Analysis of Knowledge Structures: A Network-Based Comparison of "Artificial Intelligence" and "The Renaissance" on Wikipedia

Mohammed Pathariya (mjpathar@iu.edu)
October 23, 2025

Abstract:

This project investigates and compares the underlying informational structures of two distinct knowledge domains on Wikipedia: the modern technical field of "Artificial intelligence" and the historical-cultural era of "The Renaissance." By creating and analyzing large-scale networks derived from Wikipedia's hyperlink structure, this report sought to determine if the nature of a topic fundamentally changes its network organization. The methodology involved a 2-level deep data crawl, construction of directed graphs, and analysis using centrality measures (In-Degree, PageRank, Betweenness) and Louvain community detection. The findings reveal two fundamentally different structures: the "Artificial intelligence" network is organized internally by its distinct sub-disciplines (technical, cultural, and philosophical), while "The Renaissance" network is organized externally by its historical context and lasting global consequences. This demonstrates that network analysis is a powerful tool for revealing the hidden frameworks of how human knowledge is organized.

Introduction and Background:

Complex subjects, whether technical or historical, are often defined not just by their individual facts but by the intricate web of connections that link them. Wikipedia, as a vast, human-curated knowledge graph, represents one of the largest informational networks in existence. The structure of this network—the way articles link to one another—reveals the underlying frameworks we use to organize, contextualize, and understand a given topic.

The goal of this project was to investigate and map these hidden informational structures. I sought to determine if the fundamental nature of a topic—specifically, a modern technical field versus a historical cultural era—would result in a measurably different network structure.

To explore this, I selected two distinct seed topics: "Artificial intelligence" and "The Renaissance". My hypothesis was that the "Artificial intelligence" network would be organized by its internal sub-disciplines (e.g., machine learning, philosophy, cultural depictions), while the "Renaissance" network would be organized by its external context (e.g., its historical predecessors, its consequences, and its relationship to other timelines). This report details the methodology used to collect and analyze these networks, presents the findings for each case study, and offers a comparative discussion on what these structures reveal about the nature of a topic.

Methodology:

To ensure reproducibility, this section details the precise tools, parameters, and design choices used in the analysis.

2.1. Data Collection and Caching

The network data was collected from the English-language Wikipedia, a data source that qualifies for the project's "novel data source" bonus.
Tool: The wikipedia-api library in Python was used to programmatically access the Wikipedia API.

Process (Hyperparameter): A 2-level deep crawl was initiated from each of the two seed topics ("Artificial intelligence" and "Renaissance"). A simple 1-level crawl was tested initially and found to be insufficient, producing a simple "star network" where all nodes linked back to the center, yielding uninteresting centrality metrics. A full 2-level crawl was computationally prohibitive.

Design Choice (Hyperparameter): As a balance, I implemented a 2-level crawl that was limited to the first 100 links found on the seed page. The script then visited each of these 100 pages to collect all of their subsequent links. This "100-link limit" parameter provided a dataset that was both computationally manageable and sufficiently interconnected to provide rich, meaningful analysis. The collected edge lists were saved to .csv files to prevent re-crawling.

2.2. Network Construction

The collected edge lists were used to construct graph models for analysis.

Tool: The NetworkX library in Python was used for all graph construction and analysis.
Model: The networks were constructed as directed graphs (DiGraph).
Design Choice: A directed graph was essential. A hyperlink from Page A to Page B represents a one-way endorsement or citation. An undirected graph would falsely assume this relationship is mutual, which is not the case and would corrupt the centrality analysis.

2.3. Analysis and Visualization Methods

A suite of network analysis techniques was employed.

Centrality Analysis: Three distinct centrality measures were calculated to identify important nodes:

1. In-Degree Centrality: The raw count of incoming links. Chosen to identify the most popular or cited articles.
2. PageRank: A more nuanced measure of influence, which accounts for the importance of the nodes that link to a given node.
3. Betweenness Centrality: Measures how often a node lies on the shortest path between other nodes. Chosen to identify "bridge" articles that connect disparate parts of the network.

Community Detection: The Louvain method (nx.community.louvain_communities) was selected as the main analysis method. This algorithm is highly efficient and effective for detecting community structures in large-scale networks (Blondel et al., 2008). It was run on an undirected version of the graph, as is standard for this algorithm.

Visualization: Tools: Matplotlib and adjustText libraries in Python.
Design Choice (Filtering): Initial visualizations of the largest communities were unreadable "hairballs" containing thousands of nodes. To create an interpretable visualization, I chose to generate filtered ego networks. This involved identifying the highest-PageRank node (the "ego") within the largest community and visualizing it along with only its top 25 most influential neighbors (as determined by their own PageRank scores).
Design Choice (Layout): The spring_layout algorithm was used with tuned parameters (k=0.8, iterations=100) to create a more spacious layout. The adjustText library was used to ensure no labels overlapped, resulting in a clean, report-ready diagram.

Case Study 1: "Artificial Intelligence" Network

3.1. Network & Centrality Findings

The 2-level crawl from the "Artificial intelligence" seed page yielded a substantial network, summarized in Table 1.

Comparative Network Statistics Metric	Artificial Intelligence	The Renaissance
Nodes	18,645	38,989
Edges	29,759	67,051
Density	0.000086	0.000044

The centrality analysis immediately revealed a key insight about the network's foundation. As shown in Table 2, the In-Degree metric was dominated by academic identifiers. This indicates that the AI topic on Wikipedia is not a casual collection of opinions but is built upon a bedrock of citable academic and formal sources.
The PageRank analysis provided a more balanced view of true influence. While identifiers remained high, the main Artificial intelligence article, key sub-fields (Machine learning, Natural language processing), and the culturally significant ChatGPT emerged as the most influential concepts. The Betweenness Centrality results confirmed Artificial intelligence as the primary hub, but also highlighted important "bridge" articles connecting the field to business (Amazon, Alphabet Inc.) and culture (2001: A Space Odyssey).

Rank	In-Degree Centrality	PageRank	Betweenness Centrality
1	Doi (identifier)	Artificial intelligence	Artificial intelligence
2	ISBN (identifier)	ISBN (identifier)	Amazon (company)

3	ISSN (identifier)	Doi (identifier)	Alphabet Inc.
4	Artificial intelligence	ISSN (identifier)	2001: A Space Odyssey
5	S2CID (identifier)	S2CID (identifier)	AI boom

Table 2: Top 5 Central Articles - "Artificial Intelligence"

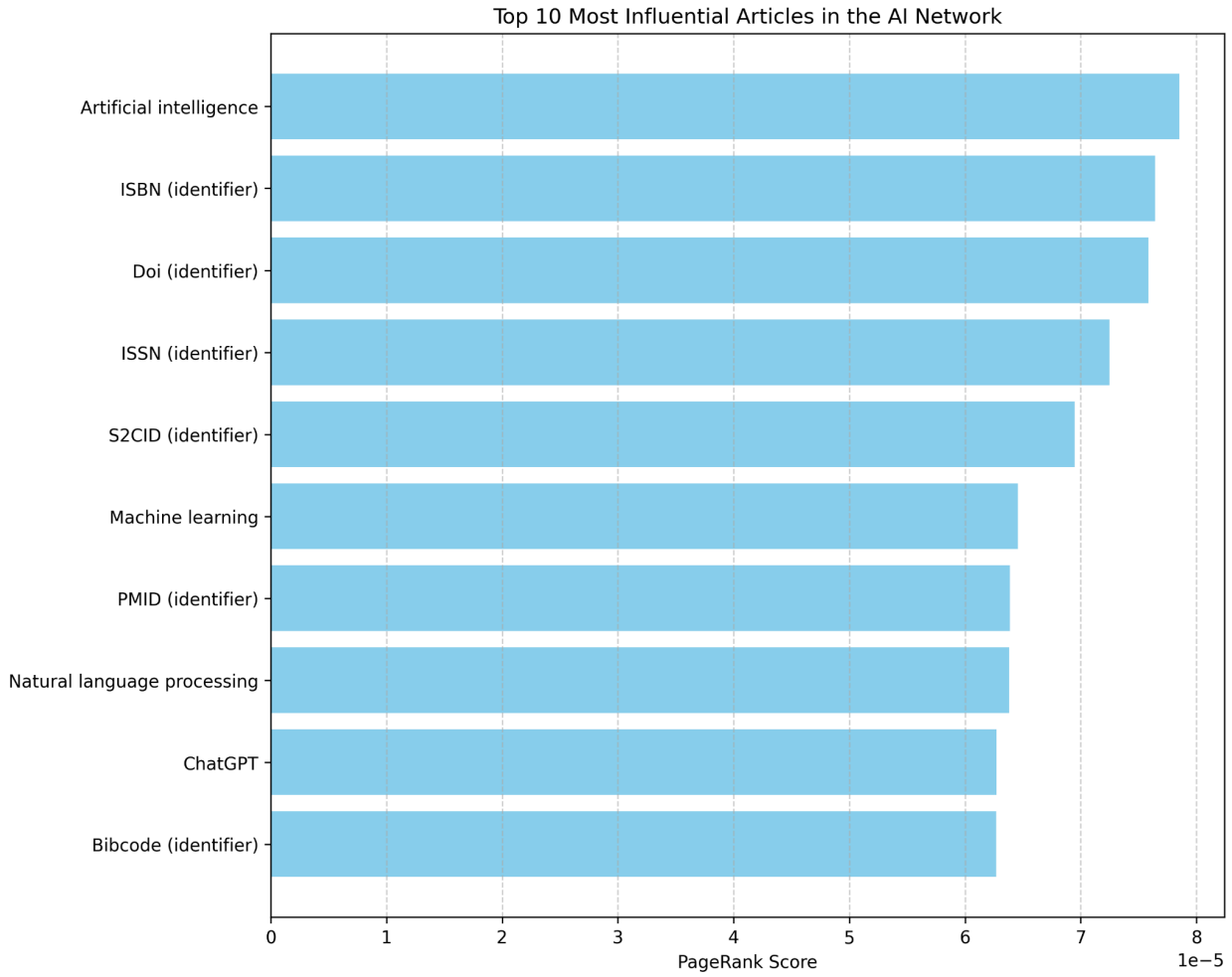


Figure 1: Top 10 Most Influential AI Articles (by PageRank)

This bar chart clearly visualizes the key conceptual and academic pillars of the AI network.

3.2. Community Structure: A Three-Part World

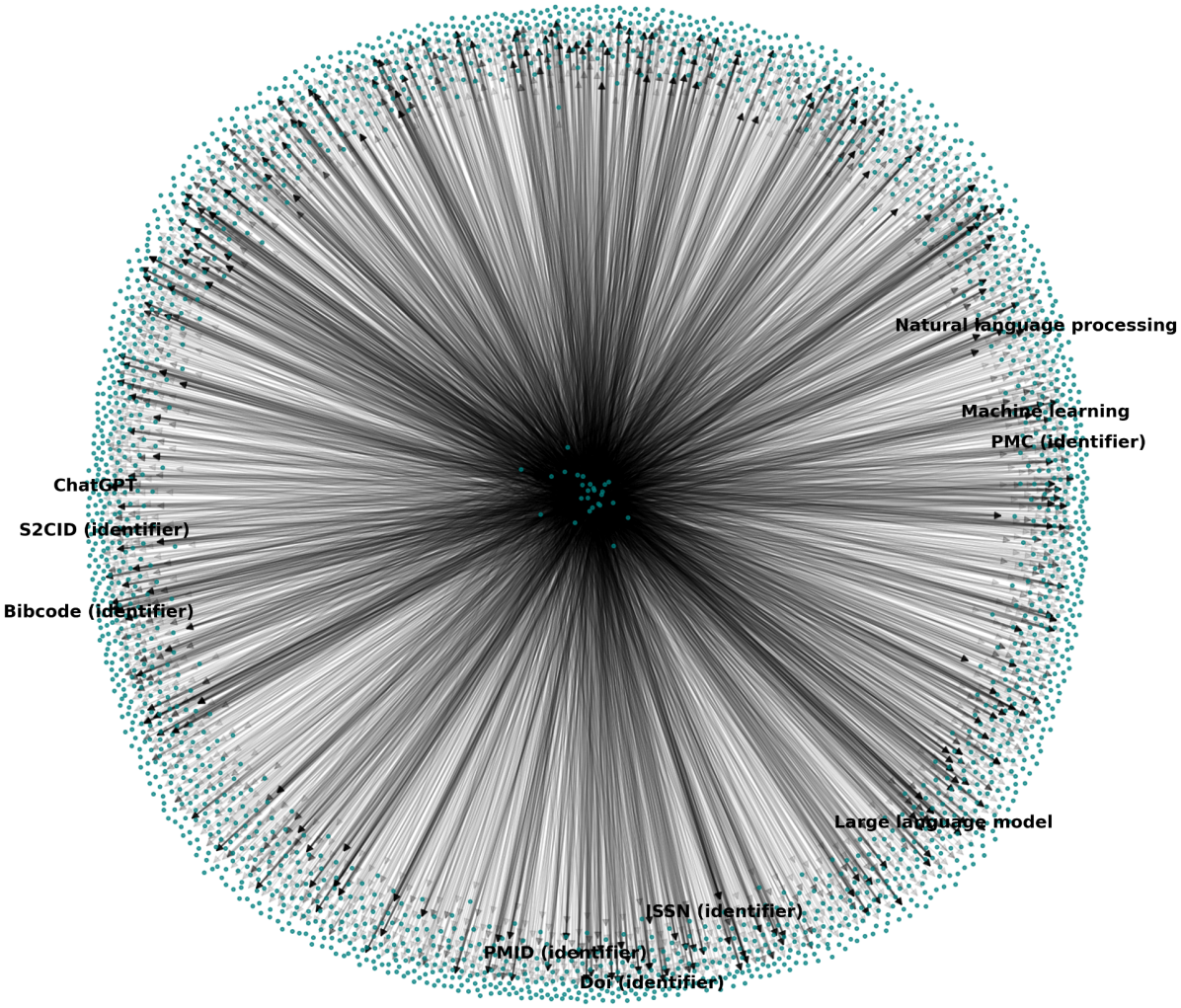
The Louvain algorithm partitioned the 18,645 nodes into 26 distinct communities. An analysis of the three largest communities (Table 3) revealed that the AI topic is not monolithic but is organized into three separate, coherent worlds.

Comm.	Size	Top Influential Nodes	Inferred Theme
1	2,308	`Doi (identifier)`, `Machine learning`, `NLP`, `ChatGPT`	The Technical & Academic Core
2	1,769	`A.I. (film)`, `2001: A Space Odyssey`, `HAL 9000`	The Cultural Imagination
3	1,764	`Ludwig Wittgenstein`, `Philosophy of mind`, `Cognitive science`	The Philosophical Foundations

Table 3: Top 3 Discovered Communities - "Artificial Intelligence"

The largest community, the Technical & Academic Core, represents the engine room of AI as a modern science. An initial attempt to visualize this 2,308-node community resulted in an unreadable "hairball" (Figure 2, Left).
To create an interpretable graph, I generated a filtered ego network (Figure 2, Right) centered on its most influential node, Doi (identifier). This visualization makes the community's structure clear: at its heart is the hub of academic citation, which directly connects foundational figures like Alan Turing, corporate entities like Amazon, modern breakthroughs like AlphaGo, and critical concepts like AI safety.

Visualization of the Largest Community in the AI Network



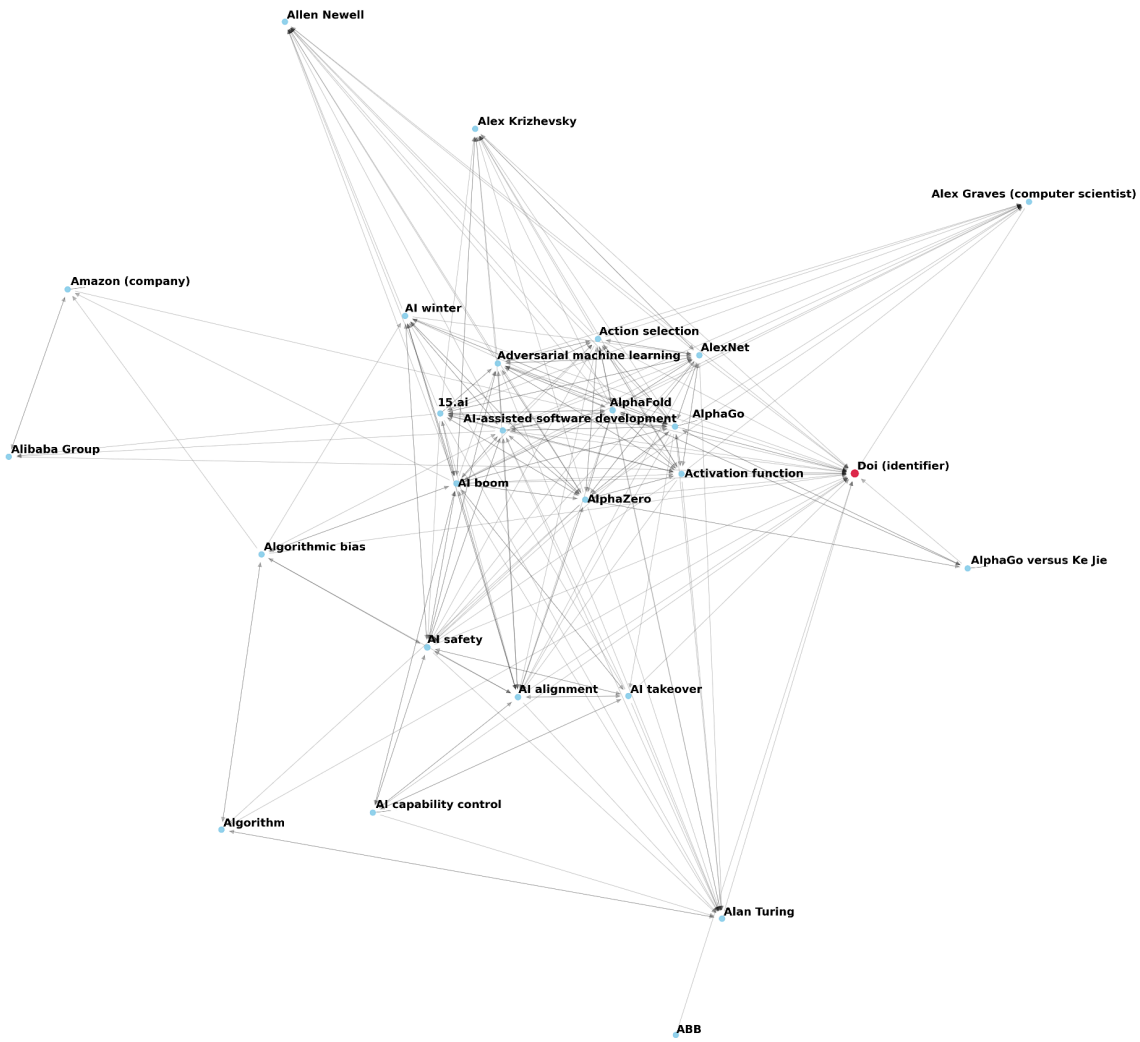


Figure 2: AI Network Visualization (Top: The full 2,308-node community, illustrating density. Bottom: The filtered ego network of Doi (identifier), revealing local structure)

Case Study 2: "The Renaissance" Network

4.1. Network & Centrality Findings

To provide a point of contrast, I applied the identical methodology to the "Renaissance" seed topic. This crawl produced a significantly larger and sparser network (see Table 1). The centrality analysis for this network told a different and more complex story. The In-Degree results were similar to AI, again dominated by academic identifiers (Table 4). However, the PageRank analysis (Figure 3) provided a genuine surprise. Alongside the usual identifiers, the articles Budapest, Buda, and Kingdom of Hungary ranked as highly influential. This suggests that within the crawled dataset, the Hungarian Renaissance was a particularly dense and well-connected sub-topic, acting as a key information hub.
The Betweenness Centrality analysis was perhaps the most insightful. Renaissance itself was the top bridge, but it was immediately followed by Age of Enlightenment, Age of Discovery, and Ancient Rome. The algorithm had perfectly positioned the Renaissance in its historical context: as the critical link between the classical world that inspired it and the modern eras it gave birth to.

Rank	In-Degree Centrality	PageRank	Betweenness Centrality
1	ISBN (identifier)	ISBN (identifier)	Renaissance
2	Doi (identifier)	Wayback Machine	Age of Enlightenment
3	ISSN (identifier)	Doi (identifier)	Age of Discovery
4	Wayback Machine	Budapest	Ancient Rome
5	JSTOR (identifier)	Buda	1948 Palestine war

Table 4: Top 5 Central Articles - "The Renaissance"

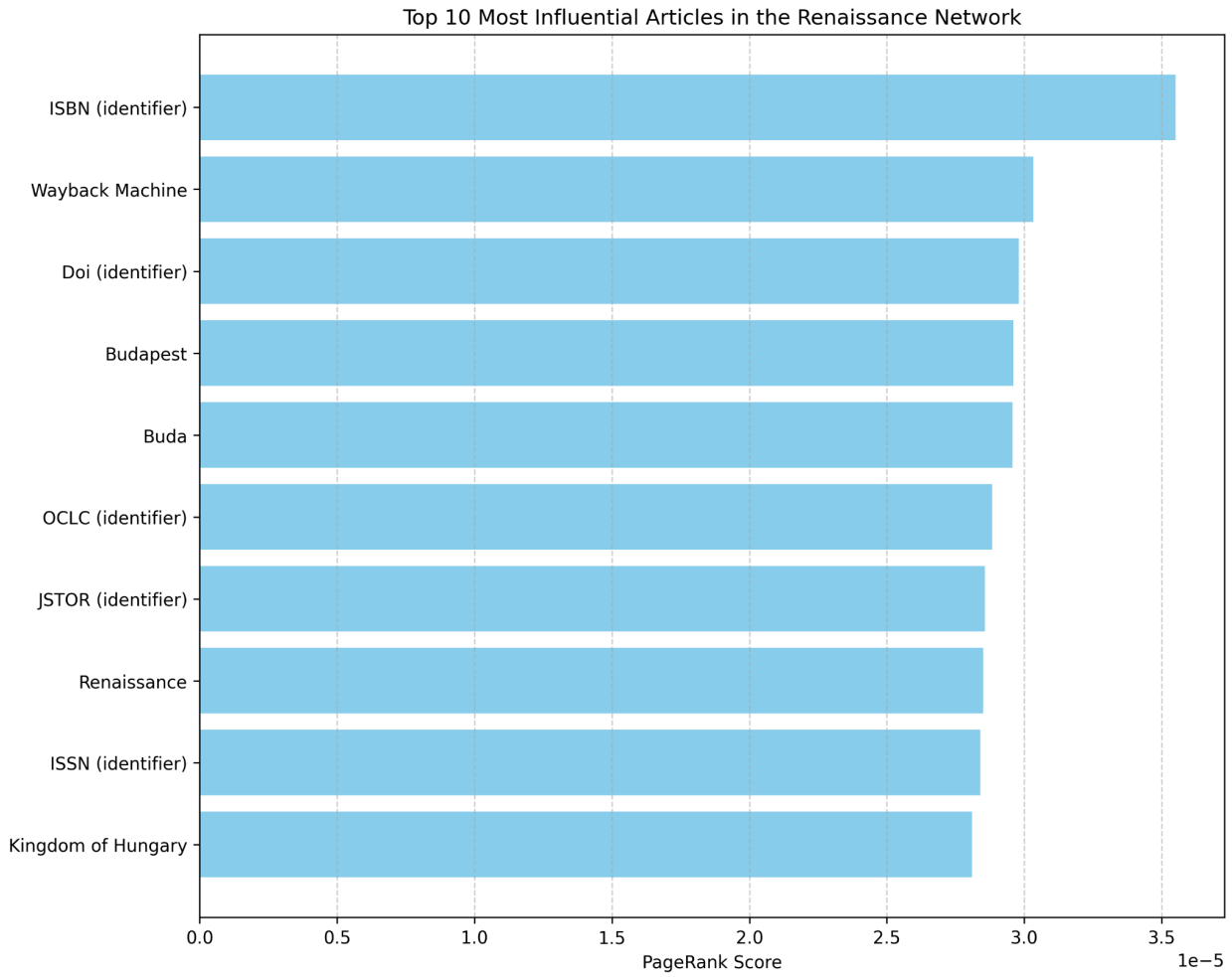


Figure 3: Top 10 Most Influential Renaissance Articles (by PageRank)

This chart shows the mix of academic identifiers and the unexpected influence of the Hungarian Renaissance topic.

4.2. Community Structure: A Network of Context

The Louvain algorithm partitioned the 38,989 nodes of the Renaissance network into 23 communities. The analysis of these communities (Table 5) revealed a structure fundamentally different from the AI network. Instead of being organized by internal sub-topics (e.g., "Italian Art," "Science"), the Renaissance network was organized by its external context and consequences.

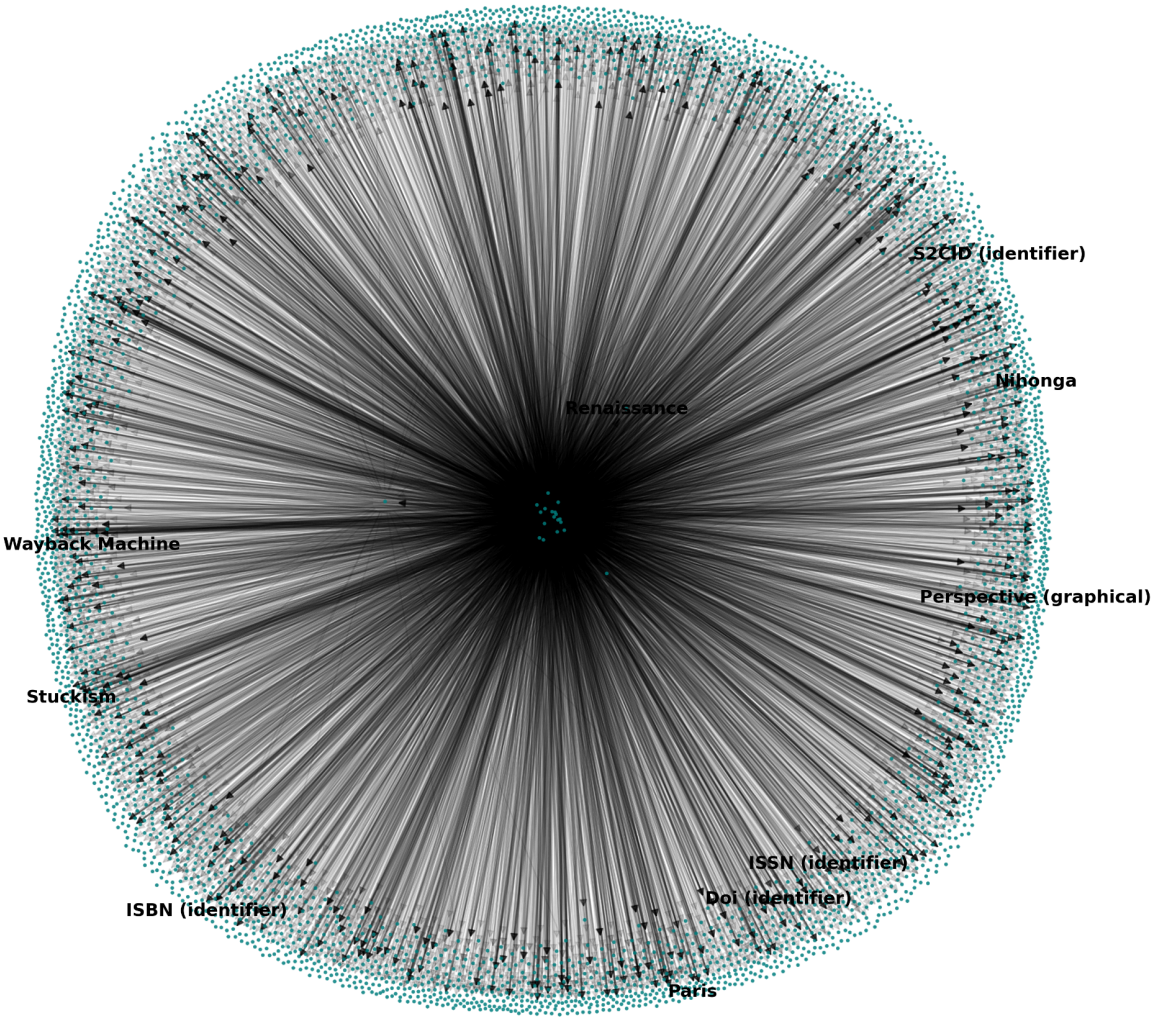
Comm.	Size	Top Influential Nodes	Inferred Theme
1	4,280	`ISBN`, `Renaissance`, `Perspective`, `Paris`, `Futurism`	The Grand Tour of Art History
2	3,838	`Age of Enlightenment`, `Cold War`, `Military history`	The European Historical Timeline
3	3,142	`Afrofuturism`, `African diaspora`, `Atlantic slave trade`	The Critical Lens of Legacy

Table 5: Top 3 Discovered Communities - "The Renaissance"

The Grand Tour of Art History community was a massive cluster connecting the Renaissance to countless other art movements, suggesting that art history is discussed as one large, interconnected conversation. The European Historical Timeline community provided the backdrop, placing the Renaissance as one point on a long timeline of major events.

Most profound was the third community: The Critical Lens of Legacy. This cluster was not about the Renaissance itself, but about its consequences. It strongly linked the "Age of Discovery" (which began during the Renaissance) to the Atlantic slave trade, colonialism, and Eurocentrism. The algorithm had uncovered a massive cluster of articles dedicated to the critical, post-colonial analysis of this European-centric era.

Visualization of the Largest Community in the Renaissance Network



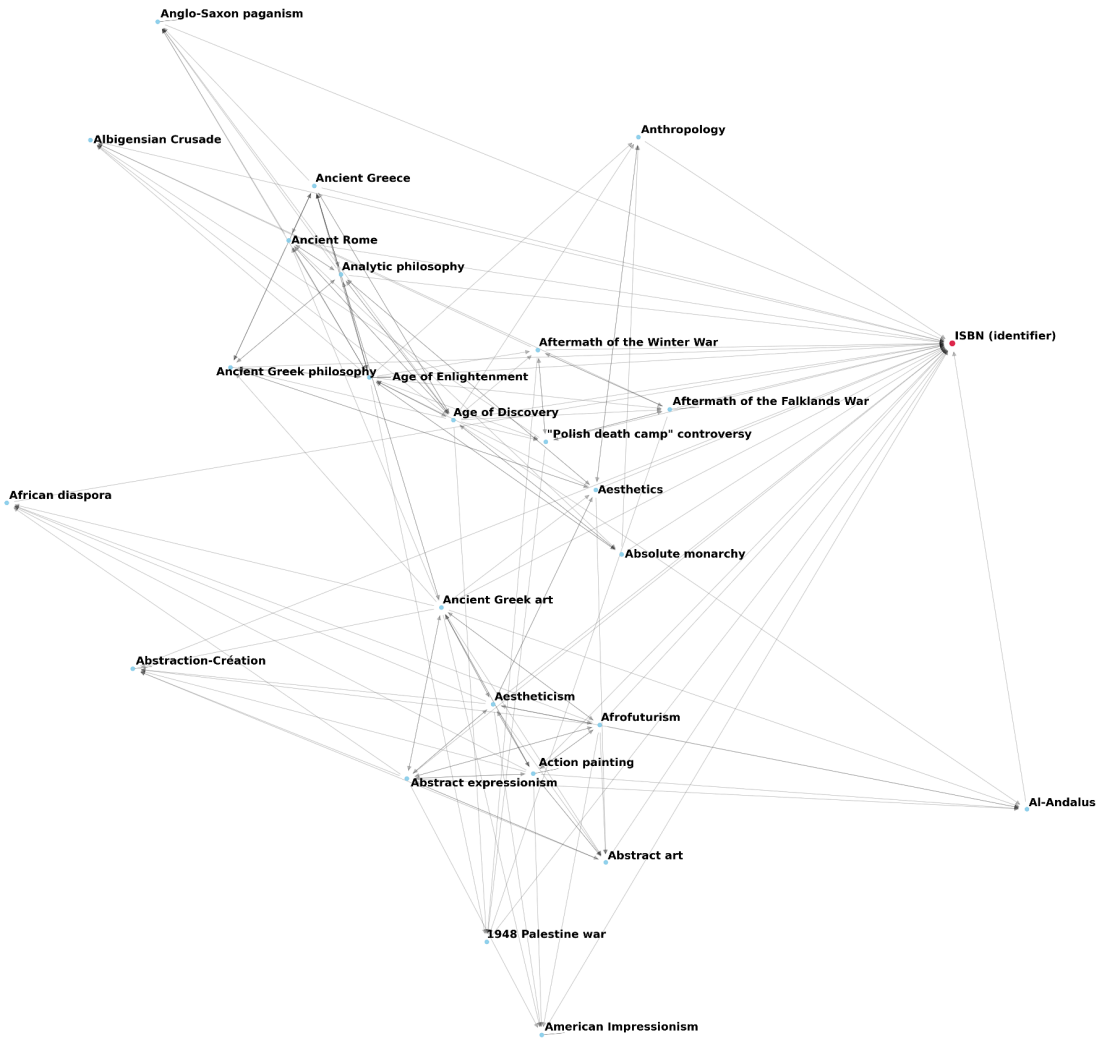


Figure 4: Renaissance Network Visualization (Top: The full 4,280-node "Art History" community. Bottom: The filtered ego network of its central node, ISBN (identifier))

The ego network reveals the web of citable art and cultural topics, from Paris and Perspective to other art movements.

Discussion & Comparative Analysis

The analysis of these two networks provides a clear answer to the initial research question. The nature of a topic fundamentally changes the structure of its knowledge network on Wikipedia.

- The "Artificial Intelligence" network is structured like a modern technical discipline: it is organized internally by its distinct sub-fields. An article belongs to the "Technical," "Cultural," or "Philosophical" community.
- The "The Renaissance" network is structured like a historical event: it is organized externally by its relationship to the rest of the world. An article belongs to the "Art History," "European Timeline," or "Critical Legacy" community.

In short, the AI network is a map of its own components, while the Renaissance network is a map of its context. This suggests that we collectively organize technical knowledge by its internal categories, but we organize historical knowledge by its place in a broader timeline and its lasting consequences.

Conclusion

This project successfully employed network analysis to map and compare the informational structures of two distinct topics on Wikipedia. By analyzing centrality and community structure, I was able to move beyond a simple list of articles to reveal the hidden "blueprints" of how we organize knowledge.

The findings show that "Artificial Intelligence" is structured as a three-part world of technical, cultural, and philosophical inquiry. In contrast, "The Renaissance" is structured not by its internal components, but by its relationship to the past (Ancient Rome), its future (The Enlightenment), and its complex global legacy (Colonialism).

This analysis demonstrates that network science is a powerful tool for the digital humanities, capable of revealing the underlying narrative frameworks we use to make sense of our world.

References

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10), P10008.

Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In Proceedings of the 7th Python in Science Conference (SciPy2008) (pp. 11-15). Pasadena, CA.

Appendix: Supplemental Materials

The full code, collected data, and supplemental materials used for this report are available on GitHub and can be used to reproduce all findings.

GitHub Repository URL:

https://github.iu.edu/mjpathar/25FA-ILS-Z639-SMM/tree/main/Assignment2