UNVEILING THE WEB OF KNOWLEDGE: INSIGHTS FROM A SIMULATED CITATION NETWORK

Esra Şekerci
Middle East Technical University
Ankara, Turkey
esra.sekerci@metu.edu.tr

ABSTRACT

This study examines a simulated citation network designed to mimic the structure and behavior of how scientific knowledge spreads. The network is made up of 150 nodes and 675 directed edges, which represent scientific papers and their citation connections. Using the Pajek software for analysis, the research delves into important structural and dynamic aspects of the network. Centrality, clustering, and main path metrics are methods used for identifying influential nodes, cohesive clusters, and primary routes of information flow. Temporal and structural patterns are probed for understanding bursts of activity and the underlying mechanisms of collaboration and knowledge dissemination. The findings offer insights into the evolution of scholarly communication and provide a framework for applying network analysis in scientific studies.

Keywords: citation network, network analysis, collaboration, structural properties, information flow

INTRODUCTION

Citations have been widely used to evaluate the impact of scientific publications, authors, and journals. Some of the citation-based metrics, such as impact factors and immediacy indices, calculate the influence of a journal, showing trends across disciplines. Citation analysis has also been applied to help identify specialties, research traditions, and paradigmatic shifts by mapping cohesive subgroups in citation networks. Network analysis techniques, such as the identification of weak components and bi-components, highlight isolated research communities and the evolution of intellectual traditions, therefore showing the structure and development of scientific knowledge.

The simulated citation network offers a valuable framework to explore the structural, dynamic, and collaborative aspects of a directed citation graph within a controlled environment. The network comprises 150 nodes (representing papers) and 675 directed edges (representing citations). It is designed to mimic certain characteristics of real-world scientific citation networks, enabling detailed analysis of knowledge dissemination within a connected system. The absence of loops and multiple lines within the network reaffirms its acyclic and non-redundant topology, consistent with the theoretical principles of directed citation graphs, where self-citations and parallel citations are inherently infrequent.

The calculated average degree of 9 signifies that, on average, each node (representing a paper) is associated with approximately nine directed connections, either as a source or target. This degree distribution highlights a moderately interconnected citation structure, indicative of a balanced network that supports knowledge flow while maintaining sparsity, a characteristic often observed in real-world academic citation systems.

Key structural properties of the simulated citation network reveal its sparse yet moderately connected nature. The largest weakly connected component (WCC), encompassing 139 nodes

(92.67% of the network), suggests a dominant backbone of connectivity while highlighting the presence of several small, isolated subgroups. In contrast, the absence of a significant strongly connected component (SCC) underscores the directed and acyclic nature of the network, consistent with the theoretical structure of citation graphs.

The network exhibits a Watts-Strogatz clustering coefficient of 0.1952 and a transitivity (global clustering coefficient) of 0.1298, indicating a moderate level of clustering with localized research communities. These metrics reflect the hierarchical structure of citations, where nodes occasionally form tightly-knit clusters or reciprocal ties, representing more cohesive areas of study.

Furthermore, the network's diameter, defined as the longest shortest path between any two nodes, is 6, spanning from Node-126 to Node-132. This moderate diameter, combined with the average degree of 9, demonstrates that the network is sufficiently connected, with most nodes reachable from others in a small number of steps. Collectively, these characteristics underline the simulated citation network's ability to model the dispersed yet interconnected nature of academic citations.

This study applies advanced network analysis techniques to address key research questions: Which nodes are most influential in the network, and what roles do they play in knowledge dissemination? How do clusters form, and how do they interact? What temporal trends and shifts characterize the network's citation activity? By leveraging tools like centrality measures, k-core decomposition, and clustering analysis, we aim to provide a comprehensive understanding of the simulated network's dynamics and its implications for scientific collaboration and information propagation.

The paper is organized as follows: Section 2 provides a literature review of the relevant literature about citation networks and their role in understanding the dissemination of scientific knowledge; Section 3 describes the methodology adopted for analyzing the simulated network and presents the results, including structural, temporal, and community-level insights; Section 4 discusses the implications of these findings for understanding networked systems of collaboration; Section 5 concludes the study and outlines future research directions.

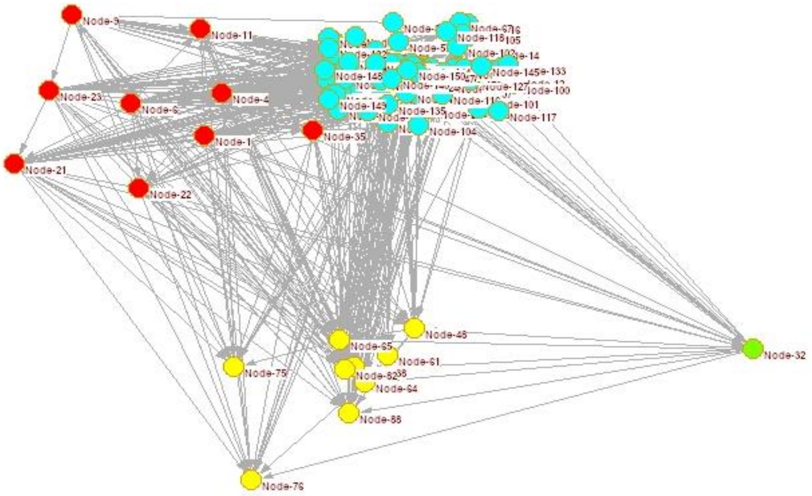
LITERATURE REVIEW

Citation network analysis provides a strong framework to trace the evolution and spread of knowledge in different scientific domains. Hummon and Dereian (1989) provided an early view into interlinkages in citation networks, using algorithms related to depth-first search to find critical links behind the development of DNA theory, showing how landmark papers can impact the trajectory of a field. Continuing this line of thinking, Liu et al. (2013) adopted citation-based analysis in mapping the evolution of Data Envelopment Analysis (DEA), pinpointing pioneering works such as Charnes et al. (1978) while simultaneously highlighting current hot subfields. Using main path analysis, Liu et al. (2014) incorporated different levels of citation relevance into the study of legal citation networks and managed to successfully identify critical legal precedents in trademark dilution cases. Lucio-Arias and Leydesdorff in 2008 provided more support for the validity of main path analysis in unraveling the codification and diffusion of scientific knowledge by tracing path-dependent dynamics of research on fullerenes and nanotubes through entropy-based indicators. Together, these studies bring out the importance of citation network analysis in knowledge diffusion dynamics underlying various fields.

METHODOLOGY

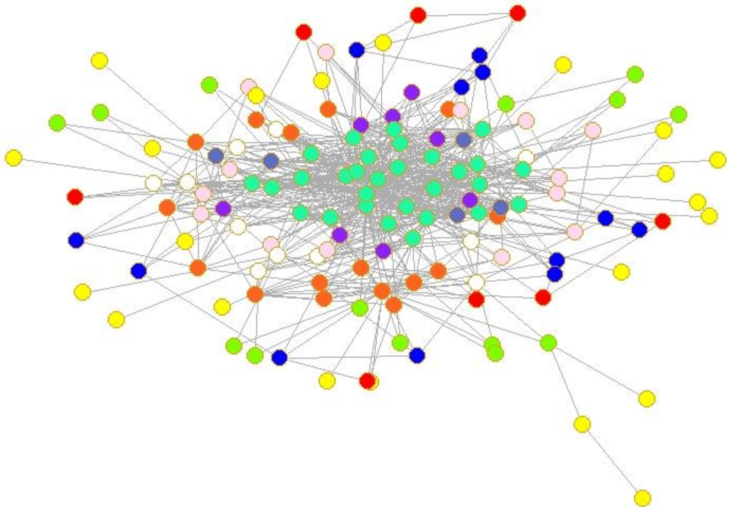
First, the acyclic nature of the citation network was verified and corrected, for conceptually, that is what citation graphs should follow: a paper may only cite previously published works; hence, this should follow a directed temporal flow from the very start of the methodology.

Figure 1. Hubs-Authorities



This visualization shows how nodes are assigned and interact with each other as hubs authorities, or both, in the citation network. Red nodes are used as hubs pointing to the crucial works. Blue nodes represent pure authorities and consist of the heavily cited articles, which, of all articles, are deemed important in the network. Yellow nodes perform multiple roles, being both hubs and authorities connecting the different parts of the network and making important additions to knowledge dissemination and integration procedures. The green node may serve as a link that directs citations to many influential works (authorities) while functioning as a primary connector in knowledge flow.

Figure 2. k-Cores in the network (without isolates)



The k-core analysis of the simulated citation network shows a hierarchical structure; the highest value of k represents the most cohesive and influential subset of articles. The highest k-core, k = 10, includes 19.33% of the nodes in this central cluster, which is considered the intellectual core of the network due to its dense interlinking through citations. Lower k-cores, such as k=0 at 7.33% and k=1 at 14.67%, are progressively less connected and peripheral articles, reflecting their reduced influence within the citation landscape. The distribution of nodes across k-cores underlines a balanced network structure, with small, tight groups complemented by a larger, loosely interconnected framework. This decomposition has identified structural cohesion but does not capture the temporal flow of knowledge. Thus, main path analysis would need to be performed as one complementary method through which the tracing of the intellectual influence evolutionary trajectory will be obtained.

Table 1. Traversal weights in the network

Line Values	Freq	Freq%	CumFreq	CumFreq%
(... 0.0000]	0	0.0000	0	0.0000
(0.0000 ... 0.0480]	607	89.9259	607	89.9259
(0.0.480 ... 0.0959]	39	5.7778	646	95.7037
(0.0959 ... 0.1439]	13	1.9259	659	97.6296
(0.1439 ... 0.1919]	2	0.2963	661	97.9259
(0.1919 ... 0.2399]	8	1.1852	669	99.1111
(0.2399 ... 0.2878]	3	0.4444	672	99.5556
(0.2878... 0.3358]	1	0.1481	673	99.7037
(0.3358 ... 0.3838]	0	0.0000	673	99.7037
(0.3838 ... 0.4317]	2	0.2963	675	100.000
Total	675	100.000		

The line values in the Citation Weights SPC (Search Path Count) analysis represent the traversal weights of the edges in the network, reflecting the significance of individual citations in terms of their contribution to the knowledge flow. The weights range from 0.00000326 (lowest) to 0.43172863 (highest). Notably, the majority of the citations (89.93%) have traversal weights of 0.0480 or less, indicating that most citations have a relatively modest influence on the overall network flow. In contrast, only a small fraction of the edges (0.296%) exhibit weights above 0.3838, identifying highly influential citations that serve as critical links in the network. This skewed distribution highlights the concentrated impact of a few key citations in shaping the flow of information. The assortativity coefficient of 0.05896 for the Citation Weights SPC indicates a slightly positive tendency for nodes to connect to others with similar traversal weights.

Table 2. Vertices with the highest traversal weights

Rank	Vertex	Value
1	148	0.8368
2	129	0.7913
3	1	0.7449
4	88	0.5378
Sum (all values)		11.7440

To identify the most important articles in the network, we analyzed the vector values derived from the Citation Weights SPC (Search Path Count) analysis. In the network, the most important article is Node-148, with a traversal weight of 0.8368, highlighting its dominant role in the knowledge flow. It is followed by Node-129 with a weight of 0.7913, and Node-1 with a weight of 0.7449, both of which serve as critical hubs in the network. Node-88, with a traversal weight of 0.5378, completes the list of the top 4 articles. All other articles in the network have traversal weights lower than **0.44**, indicating their relatively limited role in facilitating the citation flow.

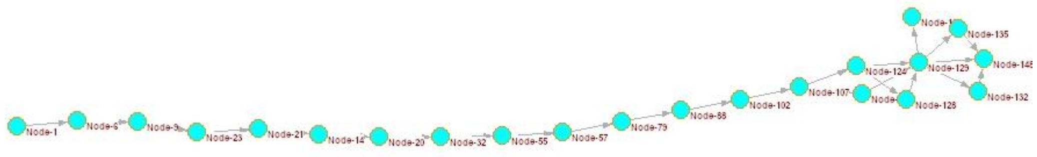
Table 3. Descriptive statistics for the traversal weights of vertices

Arithmetic mean:	0.0783
Median:	0.0038
Standard deviation:	0.1505
2.5% Quantile:	0.0000
5.0% Quantile:	0.0000
95.0% Quantile:	0.4013
97.5% Quantile:	0.4729

The statistical analysis further reinforces the skewed nature of the citation network. The arithmetic mean traversal weight is 0.0783, but the median is significantly lower at 0.0038, indicating that the majority of nodes and edges hold minimal influence, while a few dominate the network's flow. The standard deviation of 0.1505 highlights substantial variability in the weights. Additionally, at the 95th percentile, the weight is 0.4013, and at the 97.5th percentile, it rises to 0.4729, demonstrating that the highest weights are concentrated in a small number of highly influential citations.

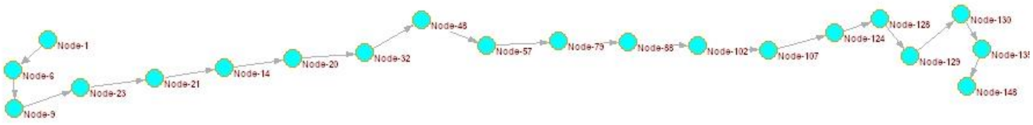
Taken together, these findings show the hierarchical and uneven structure of the citation network: most citations and papers contribute modestly, while a small fraction has an outsized impact on the network's flow. This is typical for citation networks, where the intellectual core consists of a few works and relationships monopolizing the scholarly exchange. Such influential nodes and edges are the most important when characterizing the dynamics and evolution of the network since they give rich insights into the structure and flow of academic discourse.

Figure 3. Forward local main path in the network



The Forward Local Main Path identifies the backbone of the citation network, tracing the most significant flow of knowledge through its key nodes. Starting with Node-1 as a foundational work (traversal weight: 0.7449) and progressing through pivotal connectors like Node-88 (0.5378) and Node-129 (0.7913), the path culminates at Node-148, the most influential article (0.8368). This sequence highlights the hierarchical structure of the network, where critical articles drive the intellectual discourse and connect the foundational and advanced stages of knowledge development.

Figure 4. Standard global main path



A basic distinction between the Standard Global Main Path and the Forward Local Main Path is related to the breadth of coverage: while the former considers a wider sequence of nodes, enriched by intermediary steps and allowing for a general view of the structure of the citation network, the latter limits its scope to highly influential nodes with higher weights in traversal, thus capturing a narrow yet more influential path. While the global path emphasizes connectivity and continuity, the local path highlights critical nodes and the strongest knowledge flow within the network.

DISCUSSION

The simulated citation network analysis allows for answering key research questions about the hierarchical and connected structure. Only a few influential nodes, such as Node-148, Node-129, and Node-1, are the critical hubs that enable knowledge dissemination and interconnect the clusters. Such nodes are located in the center of localized communities, integrating them into wider scholarly communication.

k-core decomposition underlines cluster formation in a layered structure, where there is a highly influential cohesive core of nodes at k=10 and peripheral contributions from lower k-cores. High-weight edges between clusters drive interactions among them, once again highlighting how influential nodes act to further knowledge.

The Forward Local Main Path shows the temporal development of knowledge, starting from the foundational works like Node-1 and moving through key connectors such as Node-88 to the most recent and influential article, Node-148. This path underlines how sequential contributions shape the network's evolution and the progress of scholarly communication.

CONCLUSION

This study shows how citation networks display a hierarchical structure in which only a few influential nodes drive knowledge dissemination and connect diverse research clusters. A k-core decomposition shows a rather cohesive intellectual core; the Forward Local Main Path detects the temporal flow of knowledge from foundational to advanced contributions. These results allow a look at the dynamics of scholarly communication and, at the same time, provide a general framework for studying citation networks. Future research might be directed at temporal trends, external influences, and changing practices that will further enrich our understanding of scientific discourse.

REFERENCES

Hummon, N. P., & Dereian, P. (1989). Connectivity in a citation network: The development of DNA theory. *Social Networks, 11*(1), 39–63. https://doi.org/10.1016/0378-8733(89)90017-8

Liu, J.S., Chen, H.-H., Ho, M.H.-C. and Li, Y.-C. (2014), Citations With Different Levels of Relevancy: Tracing the Main Paths of Legal Opinions. J Assn Inf Sci Tec, 65: 2479-2488. https://doi.org/10.1002/asi.23135

Liu, J. S., Lu, L. Y. Y., Lu, W.-M., & Lin, B. J. Y. (2013). Data envelopment analysis 1978–2010: A citation-based literature survey. *Omega, 41*(1), 3–15. https://doi.org/10.1016/j.omega.2010.12.006

Lucio-Arias, D. and Leydesdorff, L. (2008), Main-path analysis and path-dependent transitions in HistCite™-based historiograms. J. Am. Soc. Inf. Sci., 59: 1948-1962. https://doi.org/10.1002/asi.20903