IMPERIAL

Innovation Pathways and Centrality Measures for Directed Acyclic Graphs: A Literature Review

Sid Diamond

06/10/2025

Supervisor:	Tim Evans
Assessor:	Kim Christensen
Project Code:	UNIV-Evans-2
Word Count:	2384

Contents

Abstract

Identifying the technologies that drive innovation often depends on expert judgment or narrow indicators such as citation counts, which can obscure how knowledge actually accumulates. This review examines two complementary approaches to analyzing citation networks represented as directed acyclic graphs (DAGs). The first approach, which includes main path analysis and criticality-based methods, traces technological progress over time by exploiting the mathematical properties of DAGs. The second focuses on centrality measures that quantify the importance of individual nodes based on their structural position. Despite their potential, systematic evaluations of centrality measures tailored to DAGs remain limited, with relevant research dispersed across ecology, social science, and bibliometrics. This review establishes a foundation for future literature assessing how these methods perform in temporally ordered networks, which in turn can inform the development of more rigorous ways to identify influential research contributions.

Introduction

Research suggests that scholars gravitate toward familiar topics, despite evidence that combining conventional foundations with unusual research combinations yields the highest likelihood of high-impact discoveries, and in turn, innovations [1]. Understanding how knowledge accumulates is therefore fundamental to optimizing an empirically suboptimal innovation process, constrained by many factors, including the above example of individual-career risk-aversion. With the expansion of global citation databases and the adoption of data-driven methods from network science and computational modelling, it has become increasingly feasible to analyse the structure and evolution of science quantitatively using citation networks [1]. Each scientific output, whether an academic paper, patent, or clinical trial, records in its bibliography the prior work upon which it builds [2]. Linking these references forms a citation network, where each document is represented as a node and each citation from a newer to an older document is represented as a directed edge pointing backward in time.

In ideal form (after resolving temporal inconsistencies in real data) citations reference only earlier work, creating no feedback cycles due to this temporal constraint. This topological property makes citation networks directed acyclic graphs (DAGs), which have useful properties for analysing innovation. Unlike cyclic graphs where paths can loop indefinitely, the longest path between any two connected nodes in a DAG is always finite and well-defined, tracing a clear chronological chain of influence [3]. This enables path analysis methods to identify critical innovation trajectories, impossible in cyclic graphs.

Two fundamental questions emerge when analysing innovation through citation networks: Which paths trace major technological trajectories, and which individual nodes play the most influential roles?

This review establishes foundations for analysing citation networks as directed acyclic graphs.

Following the mathematical background in Section 1, Section 2 examines methods for identifying innovation pathways in DAGs. Section 3 reviews centrality measures (standard metrics for quantifying node importance) and examines challenges adapting them to DAGs, where temporal ordering constraints differ fundamentally from the undirected networks on which most centrality theory is developed.

Section 1: Mathematical Background

A graph $G = (V, E)$ consists of a vertex set V and an edge set E [4]. In directed graphs, edges are ordered pairs; $(u, v) \neq (v, u)$. In undirected graphs, edges are unordered; $(u, v) = (v, u)$. A simple graph contains no self-loops, that is, for all $(u, v) \in E$, $u \neq v$ and at most one edge between any pair of vertices.

If vertices $u, v \in V$ are connected by an edge, they are called neighbours or adjacent vertices. The neighbourhood of a vertex $v \in V$ is $N(v) = \{u \in V : \{u, v\} \in E\}$, the set of all vertices adjacent to v. The closed neighbourhood is $N[v] = N(v) \cup \{v\}$ [5].

Graph structure is encoded in a so-called adjacency matrix A, where $N = |V|$. Labelling each vertex with a unique index $i(u) \in \{1, 2, \ldots, N\}$, the matrix entries are defined as:

$$A_{i(u),i(v)} = \begin{cases} 1 & \text{if } (u, v) \in E \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

giving $A \in \{0, 1\}^{N \times N}$ for unweighted graphs. Weighted graphs use real-valued adjacency matrices not constrained to binary values.

For vertex u, in-degree and out-degree are defined as:

$$k_{\text{in}}(u) = \sum_j A_{ji}, \quad k_{\text{out}}(u) = \sum_j A_{ij} \tag{2}$$

In citation networks, in-degree equals citation count [6].

A path of length ℓ is a sequence $(v_0, v_1, \ldots, v_\ell)$ where consecutive vertices satisfy $(v_i, v_{i+1}) \in E$. The shortest path distance $d(u, v)$ quantifies the most direct route between nodes and is fundamental to many centrality measures (Section 4). Shortest paths can be computed via various algorithms, with Dijkstra's algorithm commonly used for single-source problems and Floyd-Warshall for all-pairs problems [6]. Unlike general directed graphs where longest paths may be undefined due to cycles, longest paths in DAGs are always well-defined and finite [4].

Section 2: Innovation Pathways for DAGs

While citation count (in-degree) provides a simple popularity metric, it fails to reveal how knowledge flows through time to produce innovation. Path-finding methods address this by tracing trajectories through citation networks to identify sequences of discoveries leading to technological outcomes. This section examines two approaches: main path analysis [7] and a criticality-based method [2].

Main Path Analysis (MPA), proposed by Hummon and Doreian (1989), maps knowledge flows by attempting to identify the most significant citation chains connecting foundational research to eventual applications [8]. The method follows a two-step procedure: first, assign each citation link a "traversal weight" quantifying its importance to overall knowledge flow; second, search along highest-weight links to construct the "main path".

Traversal weight quantifies how frequently a citation link participates in knowledge flows through the network. Hummon and Doreian (1989) [8] introduced three schemes: search path link count (SPLC), search path node pair (SPNP), and node pair projection count (NPPC). Batagelj (2003) later proposed search path count (SPC) alongside efficient algorithms for large-scale analysis [9]. NPPC requires $O(N^2)$ computation time, making it impractical for networks with thousands of nodes [9]. This review therefore focuses on the three remaining schemes which differ in how they count paths:

- **SPC (Search Path Count)**: counts how many times a link is traversed across all paths from sources to sinks.
- **SPLC (Search Path Link Count)**: like SPC but also counts paths originating from intermediate nodes to sinks.
- **SPNP (Search Path Node Pair)**: counts paths from all ancestors (including intermediates) to all descendants.

Unlike SPC, which treats intermediate documents as passive conduits, SPLC recognizes that intermediate papers actively generate knowledge rather than merely relaying existing ideas, whilst SPNP overcompensates by treating intermediates as comprehensive knowledge repositories [7].

Before expressing citation data as a DAG, we must address edges that create cycles. Temporal inconsistencies arise when patents cite papers published after their submission or when preprints are updated post-publication. These dating errors create forward-pointing citations and feedback loops where documents appear to cite each other cyclically (A→B→A), as demonstrated in Figure 1. Liu et al [7] recommend three solutions for these feedback loops: delete one link in the cycle, add preprints to create an acyclic structure, or group cyclic documents into a "family" node.

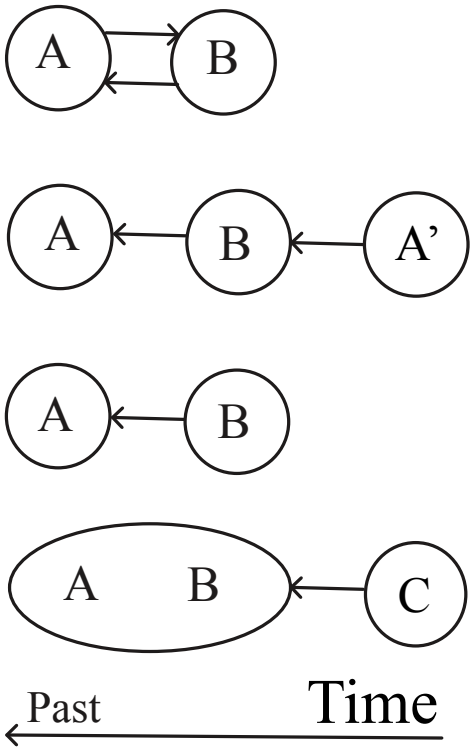

Figure 1.1. Two nodes with a temporal inconsistency forming a cycle (A→B→A), violating the DAG acyclic property.

Figure 1.2. Resolution via preprint insertion. A new node A' represents the later preprint version, creating the acyclic path A→B→A'.

Figure 1.3. Resolution via edge removal. The forward-pointing edge is deleted to break the cycle.

Figure 1.4. Resolution via node grouping. Nodes A and B merge into a single "family" node inheriting all external connections (e.g., from node C).

Figure 1: *Methods for resolving cycles in citation networks. Time flows backwards from right to left, with citations pointing backward to earlier publications.*

Traditional MPA [7] identifies a single dominant citation path through a network. Extensions such as the "network of main paths" approach allow for more holistic analysis by merging multiple local or global main paths [10]. However, Ho et al [2] present a fundamentally different approach to calculating path importance, establishing a mathematical link between technological evolution and complex networks by identifying innovation bottlenecks. Drawing from critical path method principles [11], they apply the equivalence that the longest network path (in dependencies) represents the shortest possible completion time, as this path cannot be parallelized.

The study represents innovation networks as multilayer directed acyclic graphs (DAGs), where nodes represent publications, patents, clinical trials, or regulatory approvals [2]. Starting from a regulatory approval as the source node, they quantify each document's proximity to these critical paths using three measures:

$$\text{Height: } h(v) = \max\{\text{distance from source to } v\} \tag{3}$$

$$\text{Depth: } d(v) = \max\{\text{distance from } v \text{ to any sink}\} \tag{4}$$

$$\text{Criticality: } c(v) = h_{\max} - h(v) - d(v) \tag{5}$$

where $h_{\max}$ is the height of the DAG. When $c(v) = 0$, node v lies on a longest path from source to sink, indicating it is on the critical dependency chain. Nodes with $c(v) > 0$ lie on paths c steps shorter than the longest path. The authors argue that examining near-critical nodes

(small c values) is essential because it captures innovations that may have been parallelizable or non-rate-limiting, preventing the false rejection of important innovations that single-path methods might miss.

The theoretical basis for using longest paths to identify bottlenecks draws on the finding that in DAGs embedded in Lorentzian spacetime geometry, longest paths approximate the geodesics—the routes of least resistance for information flow [12].

The paper validates this approach using the Moderna Spikevax mRNA vaccine network. Documents on longest paths ($c = 0$) included known critical developments. Documents cited in expert literature reviews [13, 14] had significantly lower criticality (median $c = 0.033$) than uncited documents (median $c = 0.169$). A Kolmogorov–Smirnov test confirmed these distributions were significantly different ($p \ll 0.0001$), demonstrating that the criticality measures for this network correlated with expert assessments of important contributions in the innovation path of the vaccine[2].

Section 3: Centrality Measures for DAGs

While path methods identify innovation trajectories through citation networks [7, 2], complementary approaches quantify individual node importance. Centrality measures rank vertices by structural position or influence [6]. The simplest metric, citation count (equivalent to in-degree [6]), has fundamental limitations that have long been recognized. Garfield [15] argued that citation frequency measures research activity and communication, not inherent significance, and is valid "only as a starting point in a qualitative appraisal." Network scientists have since developed multiple centrality indices, each capturing different notions of importance.

Degree centrality $c_D(v) = |N(v)|$ counts citations received [16], but captures only immediate connections, missing indirect influence. As discussed above and now formalized, below we highlight other well-known centrality measures [17]:

Closeness centrality measures average distance to all nodes: Closeness centrality measures average distance to all nodes:

$$c_C(v) = \left(\sum_{t \in V} \text{dist}(v, t) \right)^{-1} \tag{6}$$

where $\text{dist}(v, t)$ is the shortest path distance between nodes v and t [18, 19]. In citation networks, high closeness indicates proximity to foundational work that influenced many subsequent developments.

Betweenness centrality identifies "gatekeepers" controlling knowledge flows between research

communities:

$$c_B(v) = \sum_{s,t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{7}$$

where σ_{st} is the number of shortest paths between s and t, and $\sigma_{st}(v)$ counts those passing through v [20].

Eigenvector centrality implements recursive importance, being cited by influential papers counts more: Eigenvector centrality implements recursive importance, where being cited by influential papers counts more:

$$c_E(v) = \frac{1}{\lambda} \sum_{w \in N(v)} c_E(w) \tag{8}$$

where $N(v)$ is the set of neighbors of v (nodes that cite v in a citation network), $c_E(w)$ is the eigenvector centrality of neighbor w, and λ is the largest eigenvalue of the adjacency matrix [21].

The wide variety of centrality measures invokes the need for a unifying theoretical framework. Schoch & Brandes [5] propose neighbourhood inclusion as a defining criterion for centrality, formalizing the intuition that if an actor has the same (and possibly more) ties, it can never be less central. They prove that standard centrality indices preserve the neighbourhood-inclusion pre-order, where vertex u is dominated by v if all of u's neighbours are also neighbours of v (denoted $N(u) \subseteq N[v]$) [5].

The well-known centrality measures discussed above are typically defined for undirected, unweighted networks with a single connected component. This raises the question: how do these measures apply to DAGs, where edge directionality fundamentally alters the notion of how these indices are defined?

One approach to measuring centrality in DAGs is to adapt Google's PageRank algorithm [22]. Chen et al. [23] applied this adaptation to the Physical Review journal's citation network (353,268 papers, 1893-2003), arguing citation count alone misses influential papers. A paper's PageRank depends on both who cites it and how many references those citing papers have, being cited by influential papers with few references contributes more. Their findings showed that PageRank successfully identifies "gems": papers considered seminal in the field yet with fewer citations than anticipated.

Furthermore, centrality measures have been studied in directed acyclic graphs representing food webs. Jordán et al. [2007] [24] compared 13 centrality indices in food webs, inherently DAGs due to unidirectional trophic energy flow, finding degree and closeness centrality highly correlated, while betweenness and information centrality better captured importance when edge weights (representing flow magnitudes) were considered.

Finally, we conclude this section by noting that coherent study of centrality measures specifically designed for DAGs remains limited, with relevant work scattered across domains such as ecology

and social science. A systematic evaluation of how centrality measures perform in DAG citation networks represents an important literature gap to be explored.

Conclusion

To conclude, understanding which technologies to combine and what science must be discovered to create innovations often relies on expert consensus or popularity metrics, obscuring how intellectual efforts accumulate into technological progress [2]. This review has examined path-finding methods and centrality measures as complementary approaches for analyzing innovation dynamics in citation networks. Main path analysis and criticality-based methods provide tools for tracing technological trajectories that in some studies have shown correspondence with expert assessments [7, 2], while centrality measures offer frameworks for quantifying node importance [17]. However, systematic study of centrality measures specifically designed for DAGs remains scattered across disparate domains [23],[24]. The foundations reviewed here motivate two research directions: testing path-finding methods using numerical models of innovation processes, and evaluating how standard centrality measures perform when adapted to temporally ordered networks. Progress on these questions could inform more evidence-based approaches to identifying and supporting transformative research.

References

[1] Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Stasa Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. Science of science. *Science*, 359(6379):eaao0185, 2018.

[2] M Ho, HCW Price, TS Evans, and E O'Sullivan. Dynamics of technology emergence in innovation networks. *Scientific Reports*, 14(1):1411, 2024.

[3] V Vasiliauskaite. *Paths and Directed Acyclic Graphs*. PhD thesis, Imperial College London, 2020.

[4] James R Clough. *Causal Structure in Networks*. PhD thesis, Imperial College London, 2017.

[5] David Schoch and Ulrik Brandes. Re-conceptualizing centrality in social networks. *European Journal of Applied Mathematics*, 27(6):971–985, 2016.

[6] Michele Coscia. *The Atlas for the Aspiring Network Scientist*. 2021.

[7] JS Liu, LYY Lu, and MHC Ho. A few notes on main path analysis. *Scientometrics*, 119(1):379–391, 2019.

[8] Norman P Hummon and Patrick Doreian. Connectivity in a citation network: The development of dna theory. *Social Networks*, 11(1):39–63, 1989.

[9] Vladimir Batagelj. Efficient algorithms for citation network analysis. *arXiv preprint cs/0309023*, 2003.

[10] Bart Verspagen. Mapping technological trajectories as patent citation networks: a study on the history of fuel cell research. *Advances in Complex Systems*, 10(1):93–115, 2007.

[11] Adedeji B Badiru. *Handbook of Industrial and Systems Engineering*. CRC Press, 2013.

[12] James R Clough and Tim S Evans. Embedding graphs in lorentzian spacetime. *PLoS One*, 12(11):e0187301, 2017.

[13] Elie Dolgin. The tangled history of mrna vaccines. *Nature*, 597(7876):318–324, 2021.

[14] Shixiong Xu, Kun Yang, Rui Li, and Lu Zhang. mrna vaccine era: Mechanisms, drug platform and clinical prospection. *International Journal of Molecular Sciences*, 21(18):6582, 2020.

[15] Eugene Garfield. Citation frequency as a measure of research activity and performance. *Essays of an Information Scientist*, 1(2):406–408, 1973.

[16] Juhani Nieminen. On the centrality in a graph. *Scandinavian Journal of Psychology*, 15(1):332–336, 1974.

[17] Andrea Landherr, Bettina Friedl, and Jan Heidemann. A critical review of centrality measures in social networks. *Business & Information Systems Engineering*, 2(6):371–385, 2010.

[18] Murray A Beauchamp. An improved index of centrality. *Behavioral Science*, 10(2):161–163, 1965.

[19] Gert Sabidussi. The centrality index of a graph. *Psychometrika*, 31(4):581–603, 1966.

[20] Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[21] Phillip Bonacich and Paulette Lloyd. Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, 23(3):191–201, 2001.

[22] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.

[23] P. Chen, H. Xie, S. Maslov, and S. Redner. Finding scientific gems with google. *Journal of Informetrics*, 1(1):8–15, 2007.

[24] Ferenc Jordan, Zsófia Benedek, and János Podani. Quantifying positional importance in food webs: A comparison of centrality indices. *Ecological Modelling*, 205(1):270–275, 2007.