Journal of Informetrics 20 (2026) 101813

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Research Paper

Modeling the impact of research data unavailability on science

Jorge Chamorro-Padial [a,*], Francisco-Javier Rodrigo-Ginés [b], Rosa Rodríguez-Sánchez [c], R. M. Gil [a], Roberto García [a]

[a] Computer Science and Engineering Department, Universitat de Lleida, Jaume II 69, Lleida, 25001, Catalonia, Spain
[b] NLP & IR Group, Universidad Nacional de Educación a Distancia, Madrid, Spain
[c] Departamento de Ciencias de la Computación e I. A., CITIC-UGR, Universidad de Granada, Granada, Andalusia, Spain

ARTICLE INFO

Keywords:
Reproducibility
Open science
Knowledge propagation
Knowledge decay
Citation networks
FAIR data

ABSTRACT

Scientific progress depends on the accessibility and reproducibility of research outputs. Unfortunately, datasets and other referenced resources in academic publications frequently become unavailable over time, limiting reproducibility and reuse.

In this work, we quantitatively analyze the potential impact of research data unavailability by applying economic, probabilistic, and network based models to scientific citation networks. Rather than measuring knowledge directly, we use citation based network metrics as proxies for the dissemination and potential reuse of scientific results, and study how the absence of data-linked resources affects impact propagation and productivity-related indicators.

We further examine the resilience of citation networks under different modeling assumptions and analyze the role of highly influential nodes, or superpropagators, in amplifying the effects of dataset loss.

Our results reveal structural dependencies on vulnerable data sources and show that the magnitude of the impact depends strongly on network position and model assumptions. These findings provide quantitative evidence of the systemic consequences of data unavailability and underline the importance of long-term data preservation and accessibility policies in scientific research.

1. Introduction

Isaac Newton's famous quote, "*If I have seen further, it is by standing upon the shoulders of giants*" (Newton & Hooke, 1675), highlights how scientific progress builds upon prior knowledge. In academic writing, this process is made explicit through literature reviews and, above all, through citations, which not only link works but also reflect their meaning and relevance (Dellinger, 2005). Since the 1960s, citations have been studied as bibliometric indicators, with seminal contributions such as bibliographic coupling (Kessler, 1963) and co-citation analysis (Marshakova-Shaikevich, 1996), later used to map the structure of science itself (Small, 1999).

Unfortunately, references that underpin scientific research may become unavailable over time. This loss can involve inaccessible datasets or links, retracted or unavailable papers, or other compromised research resources. In this article, we study the consequences of such losses by modeling their effects on scientific citation networks using approaches from economics, survival analysis, and network science. Rather than assessing knowledge directly, our goal is to estimate how the unavailability of research resources affects the propagation of scientific impact and reuse within citation structures.

* Corresponding author.
E-mail address: jorge.chamorro@udl.cat (J. Chamorro-Padial).

https://doi.org/10.1016/j.joi.2026.101813
Received 22 October 2025; Received in revised form 14 March 2026; Accepted 17 April 2026
Available online 28 April 2026

Table 1
Summary of the percentage of lost references in the academic literature, based on works analyzed by Teixeira Da Silva and Nazarovets (2023).

Percentage of lost references	Reference	Year
18%-53% (after corrections: 3%)	(Lawrence et al., 2001)	2001
19%	(Wren, 2004)	2004
~50%	(Ho, 2006)	2006
14.8%-22.1%	(Wren et al., 2006)	2006
21.9% (1999) → 43.2% (2004)	(Carnevale & Aronsky, 2007)	2007
78% (2000), 56% (2003), 45% (2005)	(Thorp & Brown, 2007)	2007
16%	(Ducut et al., 2008)	2008
14.6%-17.9%	(Falagas et al., 2008)	2008
49.3%	(Wagner et al., 2009)	2009
34%-80% (1997) → 13%-22% (2012)	(Klein et al., 2014)	2014
31.2%-26.1%	(Zittrain et al., 2014)	2014
44.1%	(Sife & Lwoga, 2017)	2017
34.0%	(O'Connor & O'Connor, 2018)	2018
~23%	(Vinay Kumar & Sampath Kumar, 2019)	2019
14.7%	(Ott, 2022)	2022
20.8%	(Yayla, 2022)	2022
23.1% (LIS journals), 15.7% (Media studies)	(Niveditha et al., 2022)	2022

1.1. FAIR guiding principles for scientific data management and stewardship

In this work, we refer to the FAIR Guiding Principles for Scientific Data Management and Stewardship (hereafter, FAIR Data Principles) as a widely adopted framework that highlights the importance of data persistence, accessibility, and reuse. In particular, prior work has emphasized the role of persistent identifiers, such as DOIs, in mitigating data loss and improving long-term accessibility (Krauskopf & Salgado, 2023). However, the present study does not aim to assess FAIR compliance directly, but rather focuses on the observable consequences of data unavailability in scientific citation networks.

1.2. Open and "accessible" science: Benefits and costs

Open data enable reuse, distribution, and reproduction of scientific work, but entail costs. For researchers, these include the effort of understanding and preparing datasets (Yarnelle, 2020), while institutions must maintain infrastructures and adapt regulations (Johns Hopkins Center for Government Excellence, 2016). In academia, open data can also improve visibility, with 47% of surveyed authors reporting higher impact for their publications (Antelman, 2004).

1.3. Lost information

This work addresses the impact and cost of lost data sources, a topic with limited academic attention. Dellavalle Dellavalle et al. (2003) showed that 3.8% of internet references were inactive three months after publication and 21% after 27 months in JAMA, while Klein et al. (2014) described "reference rot" (disappearance or modification of content) and proposed a classification of articles as immune, healthy, or infected. Both studies stressed the use of persistent identifiers (DOI, URN, PURL) and archiving tools such as WebCitation, Archive.is (Costa et al., 2017), and the Wayback Machine. Table 1 summarizes different academic works that measured the percentage of lost references.

Software faces similar risks: (Escamilla et al., 2022) identified 253,590 references to repositories like GitHub in 2.66 million articles, highlighting preservation issues. Initiatives such as Software Heritage[1] aim to ensure long-term archiving (Di Cosmo, 2022).

Special attention is due to the work published by Lendvai and Sasvári (2025), where the effects of retractions on article citations are mentioned. According to this work, the majority of retracted articles have a low impact on the citation chain, but there is still a group of retracted papers that are highly cited. Schmidt (Schmidt, 2024) also notes that retracted papers may continue to be cited after retraction. Importantly, the presence of a citation does not necessarily indicate endorsement, as citations may also reflect critical engagement or dissent.

Sampath Kumar Sampath Kumar and Vinay Kumar (2013) present a closely related line of work focusing on the identification of lost web references and their recovery using archival services such as the Internet Archive's Wayback Machine. More recently, JOHN et al. (2024) analyzed the recovery of missing references in articles published by the *African Journal of Library, Archives and Information Science*.

Ultimately, the consequences of information loss are far from trivial. It undermines reproducibility, may invalidate results, and can trigger cascading effects across citation networks-confirming Garfield's early observation that a failure at one node can compromise entire branches of scientific knowledge (Garfield, 1964).

[1] https://stories.softwareheritage.org/, accessed on 4 February 2025

Table 2
Summary of the models used to measure the impact of research data losses. E = Economics, SI = Structural Influence, P = Probabilistic.

Method	Key	What does it analyse?	Analysis	References
Opportunity cost	OC	Cost of not being able to reuse lost information	E	(Buchanan, 2018)
Slow–Swam	SS	Loss of knowledge productivity due to unavailable information	E	(Kulikov, 2019)
Gozinto	Gz	Propagation based on citation paths	SI	(Rousseau, 1987)
Forward Path Search (FPSC)	FPSC	Propagation based on citation paths	SI	(Jiang & Zhuge, 2019)
Susceptible-Alert-Infected-Susceptible	SAIS	Propagation on the whole citation network	SI	(Funk et al., 2009)
Affectation & Transmission	AT	Affectation of a paper and impact on citation paths	SI	Own work
Sensitivity & Specificity	SensP	Probability of not generating new knowledge because of information loss	P	(Abby, 1994)

While these studies document the prevalence of research data loss and its immediate consequences, fewer works attempt to quantify how such losses propagate through citation networks or to estimate their broader systemic impact using formal models summarized in Table 2.

Research data loss can be generated by the unavailability of different resources (papers, datasets, links, journals, software…). In this work, we focus our empirical analysis primarily on datasets and URLs referenced by scientific articles. We studied specifically datasets and URLs because they can be extracted at scale from large corpora of academic articles to analyze. In future works, we can analyze other type of research data like software or citations. Nevertheless, the proposed framework is formulated at an abstract level and can, in principle, be extended to other types of referenced resources.

1.4. Motivation

This paper is motivated by the prevalence of research datasets and other resources that become unavailable over time, limiting reproducibility, reuse, and cumulative knowledge building. While frameworks such as the FAIR Data Principles provide normative guidance for data management, there is still limited empirical understanding of the downstream consequences of data unavailability once research outputs become embedded in the scholarly record.

In particular, the impact of dataset loss is not confined to the original publication that referenced the data. When unavailable resources are cited by influential articles, their absence may affect subsequent work through citation chains, leading to the loss of potential citations, reduced research productivity, and structural distortions in scientific communication. These effects are inherently networked and cannot be captured by analyses that focus solely on individual publications or isolated datasets.

For this reason, we adopt a network-based perspective, modeling scientific literature as a citation network in which articles differ in their structural importance and capacity to propagate impact. This approach allows us to distinguish between articles that are highly affected by data loss and those that act as superpropagators, amplifying the consequences of unavailability across multiple citation paths. By explicitly accounting for indirect effects and network position, we move beyond local assessments of reproducibility toward a systemic view of knowledge loss.

Complementing this structural analysis, we draw on concepts from economics and probabilistic modeling to quantify the opportunity costs and productivity losses associated with unavailable datasets. Economic models provide a natural framework to interpret lost citations and unrealized reuse as foregone outputs, while probabilistic approaches allow us to explore uncertainty, sensitivity, and error trade-offs in knowledge generation under data loss. Together, these perspectives enable a quantitative assessment of how dataset unavailability affects both the structure and the efficiency of scientific research systems.

Our work does not attempt to measure knowledge itself. Instead, we adopt citation-based network metrics as proxies for the dissemination and potential reuse of scientific results. Accordingly, references to *impact loss* or *productivity-related indicators* throughout this article should be understood as changes in these citation-based proxies rather than as direct measurements of epistemic value or scientific merit.

2. Methods

2.1. Scope of comparison

All models are applied to the same citation networks (arXiv and OpenAlex) to allow direct comparison. Differences in results, therefore, reflect modeling assumptions rather than network structure.

2.1.1. Model interpretation and assumptions

The analysis presented in this work relies on a set of complementary models drawn from economics, probability theory, and network science. These models differ in their assumptions, levels of abstraction, and intended interpretation. Rather than providing convergent estimates of a single latent quantity, they are used to explore how different modeling perspectives capture the structural consequences of dataset unavailability under varying assumptions.

All models in this study operate on citation-based impact proxies, as defined in Section 1.4, and therefore aim to characterize patterns of impact propagation, reachability, and sensitivity to data loss within citation networks. None of the models are intended to represent the epistemic value of individual articles or datasets, nor to provide causal estimates of knowledge production.

Economic and probabilistic models. Economic models, such as opportunity cost and productivity-related formulations, translate structural impact proxies into interpretable cost-related indicators. These models do not estimate actual monetary losses, but provide relative measures of potential productivity reduction under data unavailability. Probabilistic approaches, including sensitivity and specificity models, further characterize uncertainty and misclassification effects associated with identifying impacted nodes.

Structural propagation models. Models such as Gozinto, Forward Path Search (FPSC), and related path-based approaches focus on the topology of citation networks. They quantify how the impact may propagate along citation paths when a node associated with unavailable data is affected. In particular, the Gozinto model assumes full transmission of impact along citation paths without attenuation. This assumption is not intended to be realistic in a behavioral sense, but rather to provide an upper-bound or worst-case estimate of potential structural reachability within the network. As such, Gozinto serves as a reference model against which more constrained or dynamic propagation mechanisms can be compared.

Dynamic diffusion models. The Susceptible-Alert-Infected-Susceptible (SAIS) model captures the spread of impact through the citation network as a dynamic process influenced by alert and recovery mechanisms. In this context, infection does not represent epistemic error, but a state in which the potential reuse of a research output is affected by upstream data unavailability. Model parameters are obtained through fitting procedures designed to explore plausible diffusion regimes rather than to reproduce observed empirical diffusion processes. Consequently, results derived from the SAIS model should be interpreted as exploratory scenarios illustrating how assumptions about diffusion dynamics influence impact propagation.

Threshold-based and affectation models. Threshold parameters, including those used in the Affectation and Transmission model, define conditions under which the unavailability of a resource is assumed to have a downstream effect. These thresholds are not meant to represent empirically calibrated cutoffs, but instead function as control parameters that enable sensitivity analysis. By varying threshold values, we assess how robust the observed propagation patterns are to different assumptions about susceptibility and transmission strength.

Overall, consistency across models in this study should be understood in qualitative rather than quantitative terms. Convergent patterns-such as the identification of highly influential nodes or the amplification of impact through specific network structures-indicate robustness of the underlying structural insights, even when absolute magnitudes differ across modeling assumptions. Model-specific limitations and parameter dependencies are discussed explicitly in Section 6.

In this article, we analyze research data loss through three complementary lenses: (i) economic proxies (opportunity cost and productivity-related formulations), (ii) structural propagation models on citation networks (path-based, dynamic diffusion, and threshold-based transmission), and (iii) probabilistic models (sensitivity and specificity) that characterize uncertainty in knowledge generation under research data loss. Table 2 summarizes the models and their intended interpretation.

2.2. Quantifying the loss of knowledge: Economic approaches adapted for academia

The absence of datasets or their non-compliance with FAIR Data Principles (Stall et al., 2019) prevents replication and reuse, resulting in lost research potential. To measure this, we adapt well-known economic models such as opportunity cost and productivity frameworks.

2.2.1. Opportunity cost of non-reproducibility

In economics, *opportunity cost* is the value of the best alternative foregone (Buchanan, 2018). In academia, it reflects the missed knowledge that would have been generated if datasets were available. We model this as:

$$\text{Lost Knowledge} = N_R \times P_O \times I_K$$

Where N_R is the number of non-reproducible references, P_O is the probability of generating new knowledge if available, and I_K is the average impact of such knowledge. For instance, with $N_R = 100$, $P_O = 0.2$, and $I_K = 10$, the estimated loss is 200 potential outcomes.

2.2.2. Loss of knowledge productivity (adapted Solow-Swan model)

The *Solow-Swan model*, originally for economic growth, can be adapted to knowledge production, where researchers are labor and datasets are intellectual capital (Kulikov, 2019). Non-reproducible datasets reduce this capital, lowering productivity:

$$\Delta K = A \times L \times \left(1 - \frac{N_{NR}}{N_T}\right)$$

Here, A is the field efficiency, L is the number of researchers, N_{NR} is the number of non-reproducible papers, and N_T is the number of total papers.

These approaches show how non-reproducible datasets not only hinder replication but also reduce the system's overall productivity.

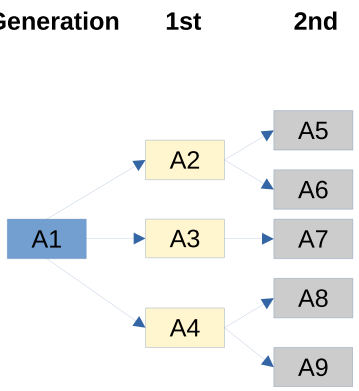

Fig. 1. Example of different generations according with the proposal of Rousseau (1987). A_5, A_6, A_7, A_8, A_9 are example of indirect cites for A_1.

2.2.3. How much does the loss of research data cost? Measuring the impact in currencies

Estimating the monetary cost of research data loss is non-trivial and highly context dependent. Costs may include recreating datasets (labor, materials, licenses, and validation) as well as downstream opportunity costs arising from disrupted citation chains. While some reports provide indicative figures for dataset creation and reuse (Ashley, 2019), there is no robust, field-independent basis for converting our citation-based impact proxies into monetary values. For this reason, we do not report currency-denominated estimates and instead focus on structurally comparable impact proxies derived from the citation network. Future work could integrate discipline-specific cost models and institutional expenditure data to support calibrated monetary assessments.

2.3. Network-Based models and the structural impact of research data lost

Traversal depth as a structural control parameter. Across the network-based models considered in this section, impact propagation is explored along citation paths up to a limited traversal depth. This depth parameter is not intended to approximate the full extent of influence propagation in real citation dynamics. Instead, it functions as a structural control parameter that balances computational feasibility, interpretability, and sensitivity analysis.

Limiting traversal depth allows the analysis to focus on early and intermediate propagation regimes, where the effects of network topology, node centrality, and model assumptions are most directly observable. As traversal depth increases, the number of reachable nodes grows rapidly, and propagation dynamics become increasingly dominated by network size rather than by localized structural features. Consequently, results obtained at finite depths should be interpreted as lower-bound estimates of potential structural impact under the modeling assumptions considered.

This unified treatment of traversal depth ensures that comparisons across models reflect differences in structural assumptions rather than artifacts of exploration range.

2.3.1. Gozinto theorem

We adopt the *Gozinto Theorem*, introduced by Rousseau (Rousseau, 1987), as a structural model to quantify how impact propagates through citation networks via both direct and indirect citation paths. Unlike metrics based solely on direct citations, the Gozinto formulation explicitly captures the cumulative influence transmitted across successive citation generations (Fig. 1).

Formally, given a citation network represented by its adjacency matrix A, the total influence matrix C is defined as

$$C = (I - A)^{-1},$$

where I denotes the identity matrix. This formulation aggregates the contribution of all citation paths of arbitrary length and can be interpreted as a Neumann series expansion.

The inverse $(I - A)^{-1}$ exists under standard assumptions for citation networks, which are acyclic by construction. In this study, the Gozinto model is used as a full-transmission propagation mechanism, providing an upper-bound estimate of potential structural reachability rather than a behavioral model of citation dynamics.

An illustrative toy example of Gozinto-based impact propagation is provided in Appendix A.

2.3.2. Forward search path

We adopt the Forward Search Path Count (FPSC) metric introduced by Jiang and Zhuge (2019) to capture the influence of articles through indirect citation paths beyond direct citation counts. FPSC is based on search path count techniques used in the analysis of scientific network evolution and incorporates a decay mechanism that reduces the contribution of more distant citation paths.

The model evaluates all citation paths connecting pairs of articles within a bounded path length, aggregating their contributions according to path distance. This maximum citation path length defines an observation window rather than a natural stopping point of influence propagation. The formal derivation of the metric is provided in Appendix B.

Table 3
Weight of each section under the IMRAD scheme, as proposed by Elrashdi and Elferjani (2024).

Section	Weight
Introduction	0.15
Methodology	0.30
Results	0.30
Discussions/Conclusions	0.25

Previous studies have shown that FPSC improves the identification of influential articles, particularly in fields characterized by dense citation interdependencies (Jiang & Zhuge, 2019). In this work, FPSC is used as a complementary path-based measure to contrast with full-transmission and threshold-based propagation models.

2.3.3. Susceptible, alert, infectious and susceptible

The Susceptible, Alert, Infectious, and Susceptible (SAIS) epidemiological model is used to analyze the spread of infectious diseases in a population (Funk et al., 2009). This model, or its variants, has been applied in network analysis to explain the transmission of research data (Sahneh & Scoglio, 2011). These models have also been employed to study the diffusion of research data in social networks, such as in the case of news propagation on the Internet (De Martino & Spina, 2015).

Drawing on this analogy, the model can be repurposed to represent how the loss of accessibility propagates through a citation network. If article B cites article A and A becomes unavailable, B is said to be infected. A third article C that cites B may, in turn, inherit this affectation. The question then arises: does C carry the same level of risk as B, or is its exposure attenuated?

Extending this reasoning, a fourth article D that cites C and simultaneously contains its own broken reference would experience a double propagation effect: one inherited via B and another originating from C itself.

Based on these dynamics, we define three conceptual states within the citation system: - Susceptible (S): articles potentially exposed to inaccessible content. - Infected (I): articles already citing unavailable or lost information. - Alert (A): articles aware of potential instability (e.g., datasets or URLs at risk), which reduces their probability of infection.

Transitions between these states can be described through differential equations analogous to those in epidemiological modeling, capturing how loss of accessibility might spread or decay over time. The formal mathematical specification of the SAIS system is presented in Appendix C, which describes the mathematical model behind SAIS.

For simplicity, in the present analysis we assume that no “healing” occurs (i.e., once a reference is lost, it remains inaccessible), corresponding to a null recovery parameter $\gamma = 0$. Future extensions could incorporate temporary recoveries or data restoration events, but these are beyond the scope of this study.

As with other network-based models in this study, propagation dynamics are evaluated within a finite exploration depth to maintain computational tractability and interpretability.

In this setting, infection states represent reduced potential for reuse rather than epistemic invalidity of the citing articles.

2.3.4. Affectation and transmission model

We propose Affectation and Transmission model. Our goal is to quantify the degree to which a scientific article is impacted by inaccessible or lost references, accounting not only for their presence but also for their structural relevance within the article. Rather than simply counting affected citations, the model estimates affectation as the proportion of the article’s weighted content that is compromised.

Each article is represented as a set of sections with different weights reflecting their contribution to scientific integrity. Missing references in sections such as methods or results are therefore treated as more critical than those in less central sections. The formal specification of the affectation model is provided in Appendix D.

The model can be extended to a transmission setting to capture how affectation propagates across citation networks. This extension employs a Linear Threshold Model to determine whether an article becomes affected through citations to compromised works. Propagation is evaluated within a bounded traversal depth, treated as a structural control parameter rather than as an empirically calibrated feature of citation dynamics.

Section weights follow the IMRAD scheme proposed by Maričić et al. (1998) and summarized in Table 3. Although necessarily approximate, this scheme provides a reasonable representation of section-level importance. More refined weighting strategies based on citation context or text analysis remain an avenue for future work (Huang et al., 2022; Thelwall, 2019).

2.4. Probabilistic impact

2.4.1. Specificity and sensitivity model

The *Specificity and Sensitivity Model*, originally proposed by Abby (1994), was designed to describe the effectiveness of the peer review process in correctly identifying valuable scientific contributions. In our context, we adapt this probabilistic framework to evaluate the ability of the scientific system to generate new knowledge in the presence-or absence-of information loss.

The formal derivation of these probabilities, based on conditional dependencies and Bayes' theorem, is presented in Appendix E. This formulation enables us to interpret information loss as a form of "noise" in the discovery process, providing a statistical framework to evaluate the robustness of research outputs in the face of non-reproducibility or data decay.

2.5. Conceptual link to economic models of knowledge production

From an economic perspective, the scientific citation network can be interpreted as a production network of knowledge, where articles act as intermediate inputs and citations encode dependency relations. In this framework, the loss of accessibility to referenced datasets constitutes a negative shock that propagates through the system, reducing the effective productivity of downstream research.

Network-based propagation models such as Gozinto, SAIS, and Affectation-Transmission are not intended as standalone economic models, but as structural mechanisms that operationalize how such shocks diffuse across interdependent production units. The resulting propagation patterns are subsequently mapped onto established economic concepts such as opportunity cost, productivity loss, and externalities-to obtain interpretable cost estimates.

3. Materials

3.1. Case study and data sources

The empirical analysis in this study focuses on the structural consequences of dataset unavailability in scientific citation networks. Given this objective, the selection of data sources is guided by the need for large-scale, well-documented citation structures and explicit links to external research resources, rather than by claims about the epistemic quality of individual publications.

Ideally, one would analyze the full scholarly record to identify and track all references to unavailable datasets. However, such an exhaustive approach is currently unfeasible due to the scale and heterogeneity of scientific literature. Instead, we adopt two complementary strategies. First, we analyze the internal citation network of arXiv to study the prevalence and structural properties of resource unavailability within citation networks. Second, we examine a set of manually curated case studies of lost datasets using OpenAlex to illustrate how the impact of dataset unavailability propagates through broader citation networks.

These data sources are therefore selected to support the analysis of citation-based impact proxies.

3.2. unarXive: Repository of links in arXiv

To construct the repository of referenced resources, we extracted all explicit URLs appearing in the full text of arXiv articles as provided by the unarXive dataset (Saier & Färber, 2020; Saier et al., 2023a,b). URLs were identified using standard pattern matching for HTTP and HTTPS links, excluding bibliographic identifiers such as DOIs, which were treated separately as formal citations within the citation network. The resulting collection includes links to a broad range of external resources, such as datasets, software repositories, project websites, and supplementary materials.

In the context of the arXiv analysis, links are treated generically as external references whose availability may degrade over time. While not all extracted URLs correspond to datasets, this approach allows us to characterize the structural embedding of external resources within the citation network and to study patterns of dependency independently of resource type. Dataset-specific analyses are introduced in subsequent sections using curated examples, where the nature of the referenced resource is explicitly verified.

The choice of arXiv as a data source is motivated by methodological rather than epistemic considerations. Our objective is not to assess scientific quality, but to analyze large-scale citation structures and the propagation of research data unavailability through them. arXiv provides full-text access, explicit URLs, and a well-defined internal citation network, which are essential for identifying external references and modeling their structural impact.

arXiv articles often contain different versions and our analysis treats each arXiv submission as a node in the citation network, focusing on its role as a citation hub rather than on its final published form. Citations are analyzed within the arXiv citation graph itself, which avoids double counting across heterogeneous sources. When OpenAlex is used, citation relationships are reconstructed independently at the work level, thereby separating preprint and journal-based analyses.

To mitigate potential biases associated with preprint literature, we complement the arXiv analysis with citation networks reconstructed from OpenAlex, which aggregates peer-reviewed journal articles across multiple publishers. The consistency of structural patterns observed across both datasets suggests that the main conclusions are not driven by idiosyncrasies of arXiv Fig. 2.

3.3. OpenAlex: Lost datasets

In addition to the large-scale analysis based on arXiv, we rely on a manually curated list of 13 datasets that are no longer accessible (see Table F.2). These datasets were originally identified in our prior systematic review (Chamorro-Padial et al., 2024), conducted following the PRISMA methodology (Page et al., 2021).

The inclusion criteria required that (i) the dataset was explicitly referenced in a peer-reviewed publication, (ii) a persistent access point (URL, repository entry, or project webpage) was reported in the publication or associated metadata, and (iii) the dataset was no longer accessible at the time of verification. Accessibility was subsequently checked on three independent dates to confirm persistent unavailability.

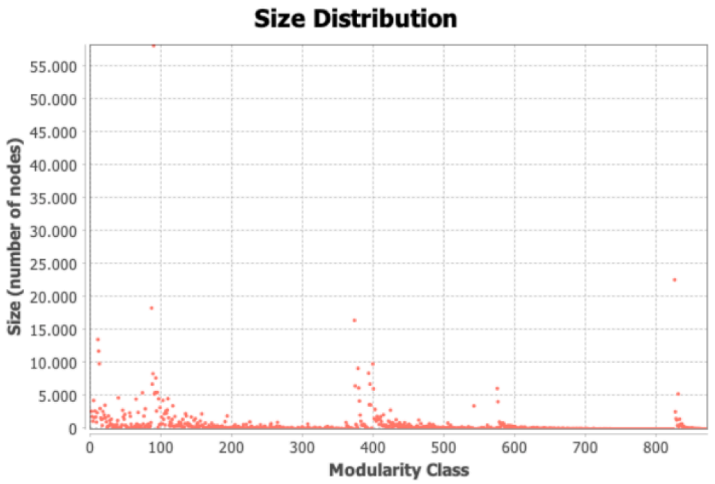

Fig. 2. Community size distribution in the arXiv citation network. Communities are detected using the Louvain algorithm and ordered by decreasing size (rank). The distribution reveals a highly modular structure, characterized by a small number of large communities and a long tail of smaller ones. Such modularity suggests that impact propagation is likely to be constrained by community boundaries rather than spreading uniformly across the entire network. The horizontal axis represents the community rank (largest to smallest), while the vertical axis indicates the number of articles per community. Algorithm: (Blondel et al., 2008), resolution adjustment: (Lambiotte et al., 2014).

Table 4
List of datasets temporarily or permanently unavailable. Access checked three times (13 March, 12 April, 14 May 2023).

Dataset name or URI	Paper	Comment	Published on
Queensland Suicide Register	(Arnautovska et al., 2015)	Temporarily inaccessible	Project webpage
National Coroners Information System	(Arnautovska et al., 2015)	Temporarily inaccessible	Project webpage
https://traps.ncipmc.org	(Rosenheim & Gratton, 2017)		Project webpage
Soybean dataset	(Rosenheim & Gratton, 2017)		Project webpage
European Pollen database	(Fyfe et al., 2015)		Project webpage
https://www.naturalearthdta.com/	(Bleasdale et al., 2020)		Project webpage
https://portal.mtt.fi/portal/page/portal/kasper/pelto/peltopalvelut/lajikekokeet	(Peltonen-Sainio et al., 2016)		Personal homepage
https://nadp.slh.wisc.edu/MDN/maps.aspx	(Yu et al., 2019)		Project webpage
https://www.pbl.nl/en/publications/2006/N2OAndNOEmissionFrom	(Philibert et al., 2012)		Project webpage
https://github.com/liuliu66/AgriPest	(Wu & Ma, 2022)		GitHub
Sen4Agrinet	(D. Sykas et al., 2022)		Project webpage
US county-level agricultural crop production typology	(Wagner et al., 2019)	Temporarily removed	FigShare

This set is not intended to be statistically representative of all unavailable datasets; rather, it provides a curated seed list to study structural propagation mechanisms in citation networks. Using a limited but verified set of unavailable datasets allows us to focus on how information loss propagates structurally through citation networks rather than on estimating the overall prevalence of dataset disappearance.

To ensure the accuracy of the automatically detected inaccessible resources, we conducted a manual verification of the six OpenAlex case studies. The detailed results of this verification are reported in Appendix J.

Reconstructing complete citation networks for all 13 datasets proved computationally and methodologically challenging. Therefore, we selected six datasets as illustrative case studies to capture a range of citation patterns and network positions (Bleasdale et al., 2020; D. Sykas et al., 2022; Peltonen-Sainio et al., 2016; Philibert et al., 2012; Rosenheim & Gratton, 2017; Wagner et al., 2019). The selected datasets span multiple scientific domains, including ecology, agriculture, geosciences, and remote sensing, and comprise a mix of observational, experimental, and derived data products. Rather than being statistically representative of all lost datasets, these cases are intended to capture structural diversity in terms of citation volume, network position, and downstream influence. This diversity enables a comparative analysis of how dataset unavailability propagates differently depending on disciplinary context and network structure. While not statistically representative of all lost datasets, these cases allow us to explore how dataset unavailability can lead to different propagation dynamics depending on network structure and article centrality Fig. 4.

Using the OpenAlex API,[2], we constructed the corresponding citation networks, comprising 662,909 articles and 662,903 edges.

This limitation is discussed further in Section 6 as a trade-off between computational feasibility and impact coverage.

[2] https://openalex.org/ accessed on 2 May 2025

Table 5
Number of descendants per depth level in arXiv citation network.

Propagation level	Affected articles
0	13,206
1	1552
2	230
3	20
4	1
5	1

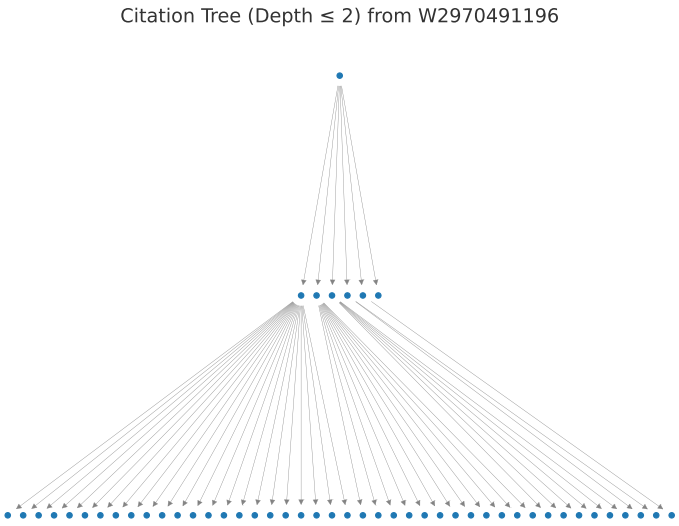

Fig. 3. Citation network induced by the unavailable dataset reported in (Wagner et al., 2019), limited to a traversal depth of two. Nodes are arranged hierarchically by citation depth, with the root representing the lost dataset and subsequent levels corresponding to direct and indirect citing articles. The figure illustrates the highly branching nature of early-stage impact propagation and the concentration of influence in a small number of intermediate nodes, motivating the use of depth-limited and sensitivity-based analyses in the network models considered.

4. Results

4.1. General results

arXiv. After an initial exploration of our dataset, we analyzed a total of 160,125 articles and identified approximately 13,206 inaccessible references.[3] Affected references are distributed through the different levels of the citation network as shown by Table 5.

OpenAlex. Working with the full citation trees for these datasets is challenging; therefore, for illustrative purposes, we show only the citation tree for the paper (Wagner et al., 2019) (OpenAlex ID: W2970491196) (see Fig. 3).

To collect the data, we performed a pre-order traversal of the citation tree. For each reference, we queried the OpenAlex API to retrieve information about its descendants. Given the huge size of the citation network, we limited the traversal to a maximum depth of 4,[4] as we consider this sufficient for the objectives of our study.

4.2. Network models: Gozinto, SAIS, and affectation–transmission

This section reports the results obtained from applying three network-based propagation models-Gozinto, SAIS, and Affectation–Transmission-to the arXiv and OpenAlex citation networks. All models are applied to identical sets of initially affected nodes and share the same underlying network structures, enabling direct comparison of propagation patterns under different modeling assumptions.

[3] See Section *Limitations* for more details regarding the limitations observed in relation to the number of lost references. For this reason, we refer to 13,206 as an approximate value.

[4] Regarding depths, it's important to distinguish between the depth of the citation trees from arXiv and OpenAlex, and the depth limit set for each model, that will depend on the specific characteristics of the model.

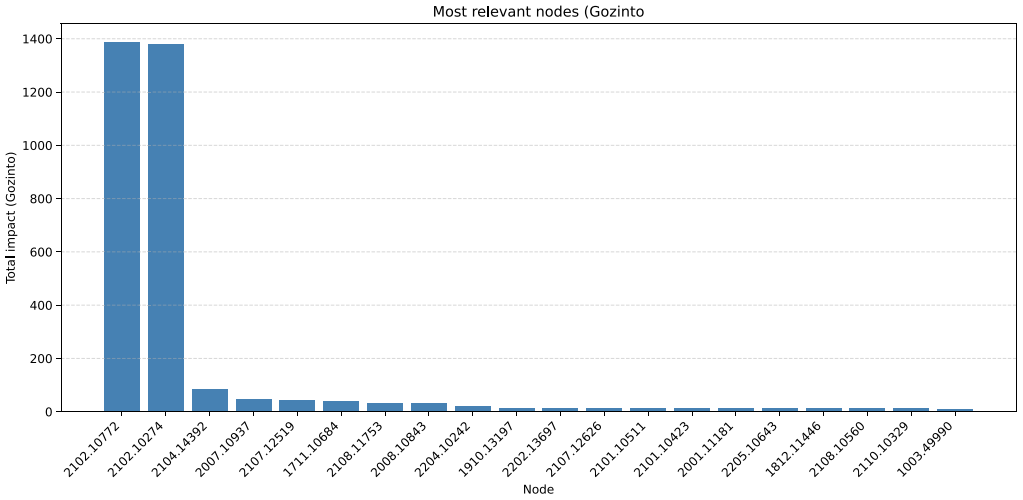

Fig. 4. Top-ranked nodes according to the Gozinto model, ordered by cumulative propagated impact. The distribution highlights a strong concentration of impact in a small number of structurally central nodes, acting as superpropagators within the citation network. This pronounced asymmetry supports the use of network-based impact proxies to identify vulnerable points where dataset unavailability can generate disproportionate downstream effects.

Table 6
Gozinto model over OpenAlex network: number of affected articles at each depth. Nodes are identified by their OpenAlex ID.

OpenAlex ID	Depth 1	Depth 2	Depth 3	Depth 4	Depth 5
W2031816812	1	83	5344	103,143	481,069
W2336803818	1	50	894	9138	53,647
W2549187017	1	35	486	2750	0
W4225728415	1	29	220	444	0
W3097366679	1	22	190	977	4076
W2970491196	1	6	44	256	0

Gozinto. The Gozinto model produces the most extensive propagation across the citation network. Impact grows rapidly with traversal depth, particularly for structurally central nodes. In the OpenAlex case studies, nodes `W2031816812` and `W2336803818` exceed 100,000 affected articles at depth level 4 (Table 6), illustrating strong amplification effects inherent to path-based propagation.

Fig. 4 shows the top-ranked nodes according to cumulative propagated impact, revealing a highly skewed distribution in which a small number of articles concentrate a large fraction of downstream effects. Fig. 6 further highlights these nodes as superpropagators, defined as articles whose unavailability leads to disproportionately large cascades through the citation network. From an economic standpoint, this asymmetry reflects the presence of strong non-linearities in knowledge production networks, where shocks to a small number of structurally critical inputs generate disproportionately large downstream productivity losses.

The relationship between propagated impact and cumulative received impact is shown in Fig. 5. The wide dispersion observed indicates that articles acting as major propagators are not necessarily those most exposed to upstream losses, reflecting distinct structural roles within the citation network.

Additional results can be found in Appendix H.

Susceptible–Alert–Infected–Susceptible (SAIS). The SAIS model yields propagation patterns that are qualitatively similar to those observed under Gozinto but with lower overall magnitudes. When recovery is not allowed ($\gamma = 0$), the number of infected and alert nodes increases monotonically and the final spread is mostly contained within depth 4. (Fig. 8). Allowing recovery leads to earlier saturation and reduced long-term exposure (Fig. 7).

Table 7 summarizes the number of infected nodes per depth for the OpenAlex case studies. While certain nodes still generate substantial downstream spread, propagation remains more constrained than under the Gozinto model, reflecting the probabilistic and state-based nature of SAIS dynamics.

Affectation–transmission. Propagation under the Affectation–Transmission model is strongly controlled by the threshold parameter. When the threshold is set to zero, affectation propagates along all reachable citation paths within the traversal depth, corresponding to an upper-bound scenario. For any positive threshold value, propagation is rapidly constrained and typically does not extend beyond depth level 3 (Fig. 9).

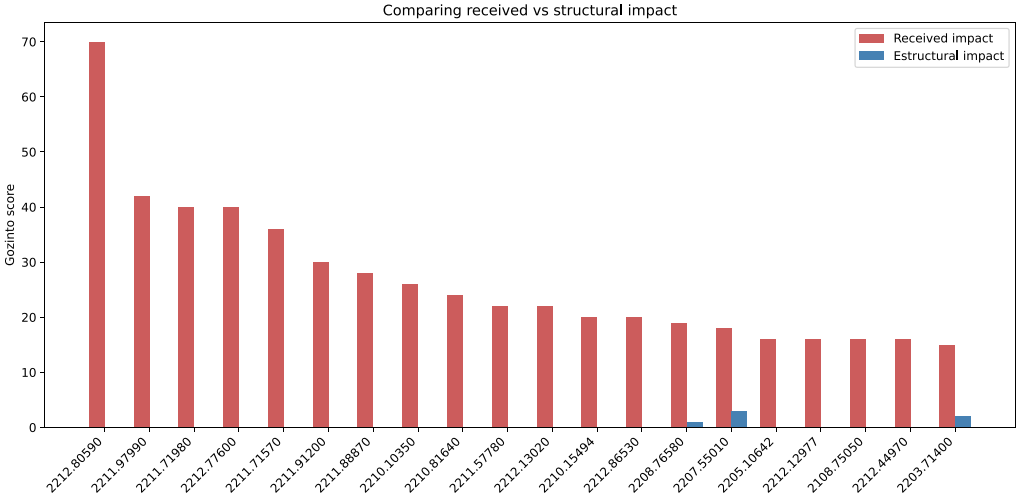

Fig. 5. Comparison between cumulative received impact and structural (propagated) impact across citation nodes, based on the Gozinto model. Nodes are ordered by decreasing received impact. The figure reveals a clear divergence between vulnerability and structural influence: nodes that receive the highest impact are not necessarily those that contribute most to impact propagation. This decoupling highlights the presence of distinct structural roles within the citation network and motivates the use of multiple complementary impact proxies. This contrast is indicative of superpropagator behavior being driven by network position rather than by exposure alone.

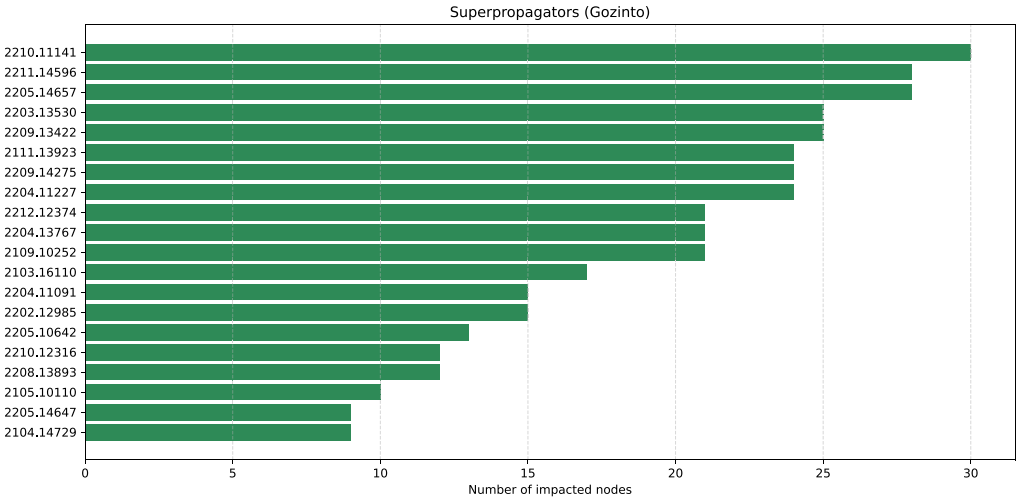

Fig. 6. Superpropagator nodes (articles) identified by the Gozinto model. Each bar represents an article in the citation network whose unavailability leads to downstream impact propagation. The horizontal axis shows the number of affected articles reached through citation paths, combining direct and indirect effects. The distribution highlights that a small number of structurally central articles generate disproportionate downstream impact, reinforcing the role of superpropagators in the propagation of dataset unavailability.

Table 7
SAIS model over OpenAlex network: number of infected nodes at each depth for every initially affected node (final step). Nodes are identified by their OpenAlex ID.

OpenAlex ID	Depth 0	Depth 1	Depth 2	Depth 3	Depth 4
W2031816812	1	46	2590	39,393	158,600
W2336803818	1	25	425	3856	17,656
W2549187017	1	18	214	914	0
W4225728415	1	13	81	142	0
W3097366679	0	9	78	399	1352
W2970491196	1	2	15	83	0

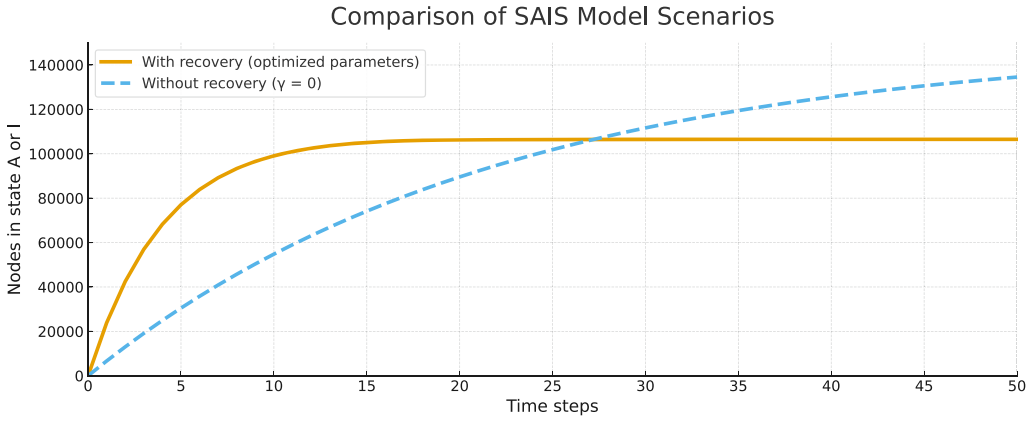

Fig. 7. Comparison of SAIS model scenarios with and without recovery. The curves show the evolution of the number of nodes in the Alert (A) state over time. The scenario without recovery ($\gamma = 0$) represents irreversible loss of accessibility, while the recovery scenario allows transitions back to the Susceptible state using parameters obtained through numerical fitting in a synthetic simulation setting. The results illustrate how recovery mechanisms can substantially reduce long-term exposure and lead to saturation, highlighting the potential mitigating role of data restoration and preservation strategies. These results are intended as a qualitative scenario comparison rather than as a calibrated prediction.

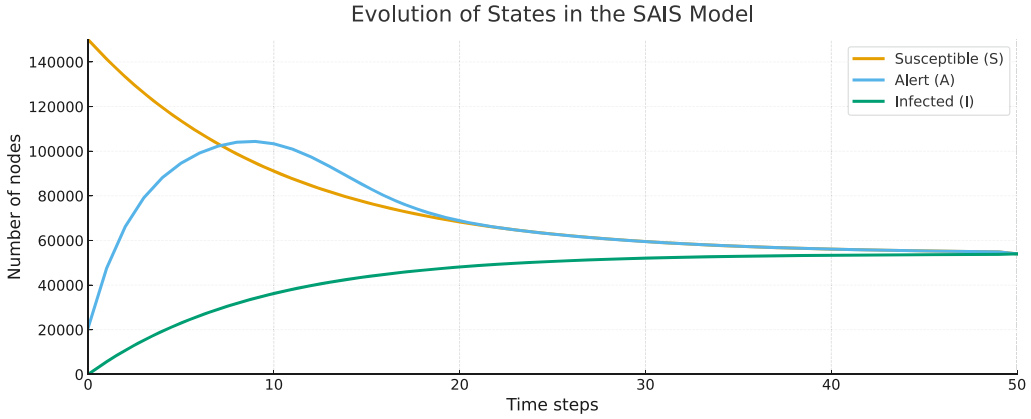

Fig. 8. Temporal evolution of Susceptible (S), Alert (A), and Infected (I) states in the SAIS model assuming no recovery ($\gamma = 0$). The simulation shows that, in the absence of recovery mechanisms, the system converges to a steady state where a substantial fraction of nodes remains permanently affected or alert. This behavior illustrates the persistence of vulnerability under irreversible information loss and provides a baseline for comparison with recovery-enabled scenarios.

Beyond threshold values of approximately 0.5, no further changes in propagation patterns are observed, indicating a saturation regime under the modeling assumptions considered.

Model comparison. Table 8 compares the number of affected nodes per depth across the three models. The results highlight substantial differences in propagation scale and depth: Gozinto produces the deepest and most extensive cascades, Affectation–Transmission yields the most restrictive propagation, and SAIS occupies an intermediate position. Fig. 10 summarizes these differences by comparing the total number of affected nodes per initial source article across models.

To assess the robustness of the propagation results with respect to the traversal depth parameter, we conducted an additional sensitivity analysis extending the exploration depth beyond the values used in the main analysis (up to depth = 6) for selected OpenAlex case studies. Increasing traversal depth in large citation networks leads to a rapid combinatorial growth in the number of reachable nodes, which significantly increases the computational cost of the propagation models in terms of both processing time and memory requirements. For this reason, the main analysis focuses on early propagation regimes where structural effects are most interpretable and computationally tractable.

The additional analysis confirms that increasing the traversal depth primarily affects the magnitude of propagation but does not alter the qualitative conclusions of the study. In particular, the highly skewed distribution of propagated impact and the identification of superpropagator nodes remain stable. Moreover, the propagation dynamics exhibit a saturation trend beyond depth levels around 5, indicating that further increases in depth mainly capture additional peripheral nodes rather than altering the structural patterns observed. Detailed values for this sensitivity analysis are reported in Appendix I.

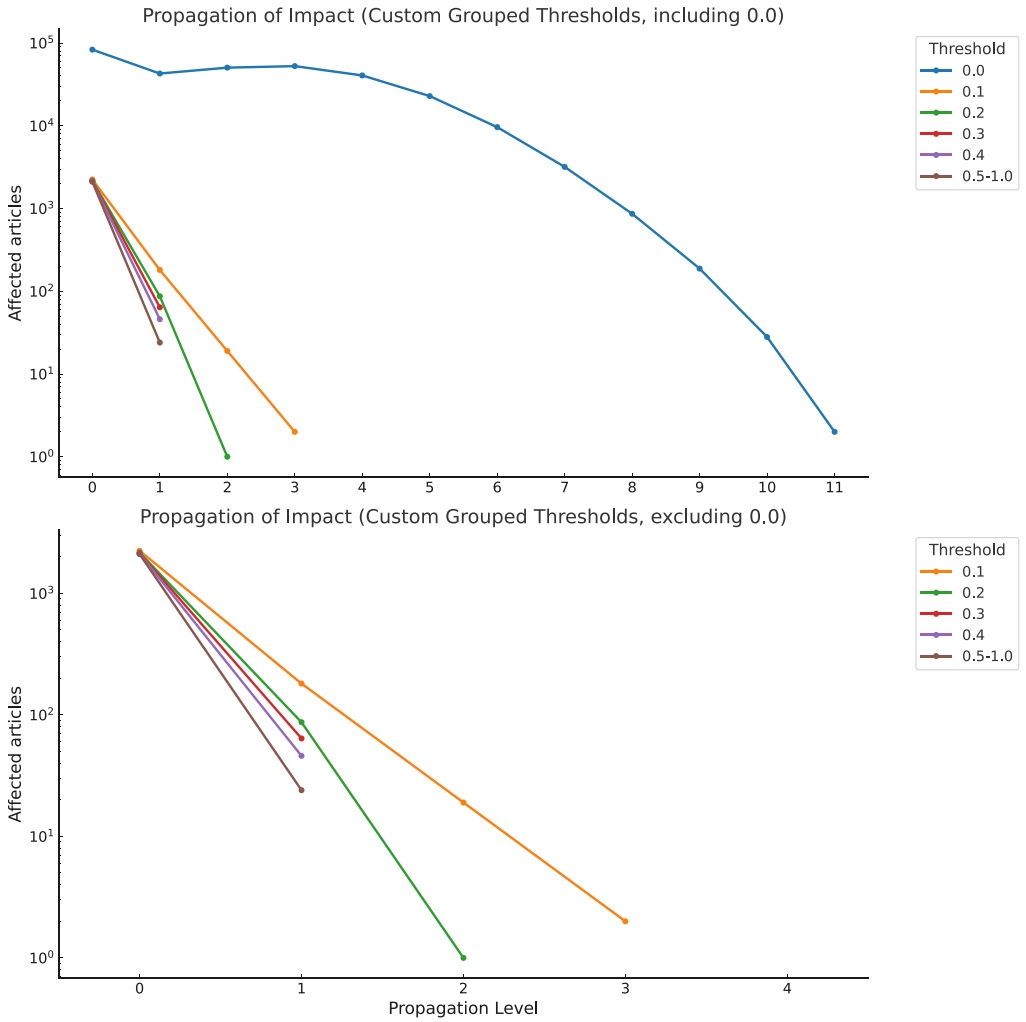

Fig. 9. Propagation impact under the Affectation-Transmission model for different threshold values. The upper panel includes the degenerate case $t = 0.0$, corresponding to unrestricted propagation, while the lower panel excludes this case to highlight differences among positive thresholds. Threshold values are treated as a structural control parameter rather than as empirically calibrated quantities. Results show that $t = 0.0$ defines a worst-case regime with unbounded propagation, whereas all $t > 0$ lead to rapidly constrained propagation, indicating robustness of the qualitative behaviour across positive threshold values.

Table 8
Comparison of affected nodes per depth across models.

Depth	Gozinto	Affectation Transmission	SAIS
0	6	6	5
1	225	225	113
2	7178	7178	3403
3	116,708	116,708	44,787
4	538,792	0	177,608

4.3. Economic cost estimations

As discussed, our aim was to estimate the cost of research data lost based on various models drawn from the field of Economics. These economic formulations should be interpreted as stylized mappings from structural propagation patterns to economic loss proxies, rather than as calibrated macroeconomic predictions. Their purpose is to translate network-level vulnerability into quantities that are meaningful within the economic literature on opportunity cost, productivity, and externalities.

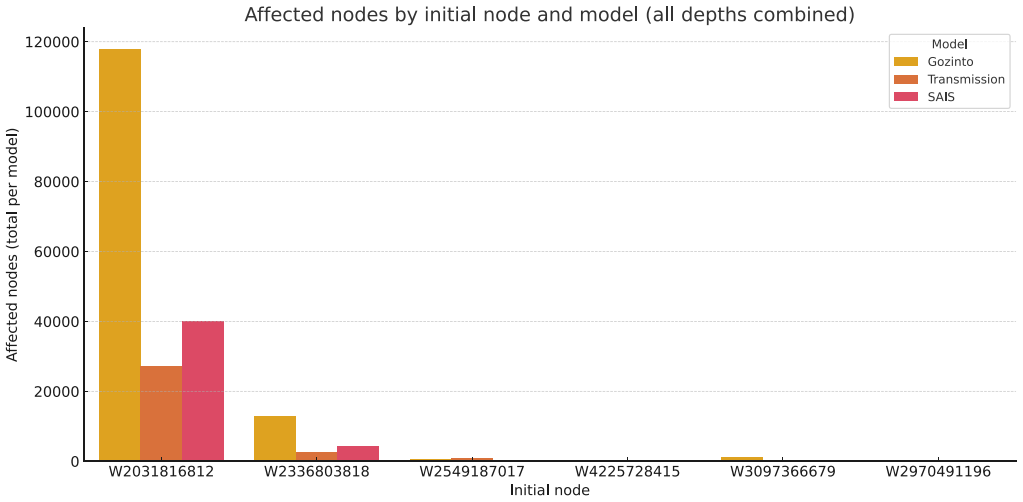

Fig. 10. Number of affected nodes per initial source article in the OpenAlex citation network, aggregated across all propagation depths. Bars correspond to different propagation models, illustrating how estimated impact depends strongly on the underlying modeling assumptions. The large variation observed across models and source nodes highlights the sensitivity of impact estimates to propagation mechanisms and supports the use of comparative, multi-model analyses rather than single-metric assessments.

Table 9
Estimated probabilities from the specificity and sensitivity model applied to the citation network.

Probability	Value	Description
$P(X_1)$	0.1945	Marginal probability of generating new knowledge.
$P(X_0)$	0.8055	Marginal probability of not generating new knowledge.
$P(X_1 \mid S_1)$	0.0826	Probability of generating new knowledge when no information loss occurs.
$P(X_0 \mid S_1)$	0.8506	Probability of not generating new knowledge when no information loss occurs.
$P(X_1 \mid S_0)$	0.0015	Probability of generating new knowledge when information loss occurs.
$P(X_0 \mid S_0)$	0.0222	Probability of not generating new knowledge when information loss occurs.

Opportunity Cost. We consider $N_R \approx 13,206$, and $P_O \approx 0.1268$, which was estimated as the average number of incoming citations an article receives in our network. Finally, $I_K \approx 0.2826$ represents the estimated average impact of articles, up to a citation depth of 2.

$$\text{Lost Knowledge} = N_R \times P_O \times I_K$$

$$473.2195 \approx 13{,}206 \times 0.1268 \times 0.2826$$

$$\text{Lost Knowledge} \approx 4,732,195 \text{ potential outcomes}$$

Loss of knowledge productivity (adapted Solow-Swan model). Our adapted formula for the Solow-Swan model is as follows:

$$\Delta K = A \times L \times \left(1 - \frac{N_{NR}}{N_T}\right)$$

Where:

- A is calculated analogously to P_O, with $A \approx 0.1268$.
- $L = 335,970$ represents the number of unique authors in the network.
- $N_{NR} \approx 13,206$
- N_T is 348,859

$\Delta K = 40988.34$ potential units of productivity lost.

4.4. Sensitivity and specificity

The *Specificity and Sensitivity Model* was applied to the citation network to estimate the probability of generating new knowledge under different information integrity conditions. The computed values are summarized in Table 9.

The results reveal a strong dependence between the availability of research data and the generation of new knowledge. When all references are accessible, approximately 8% of articles are likely to produce novel contributions. However, when one or more references become unavailable, this probability drops to less than 0.2%.

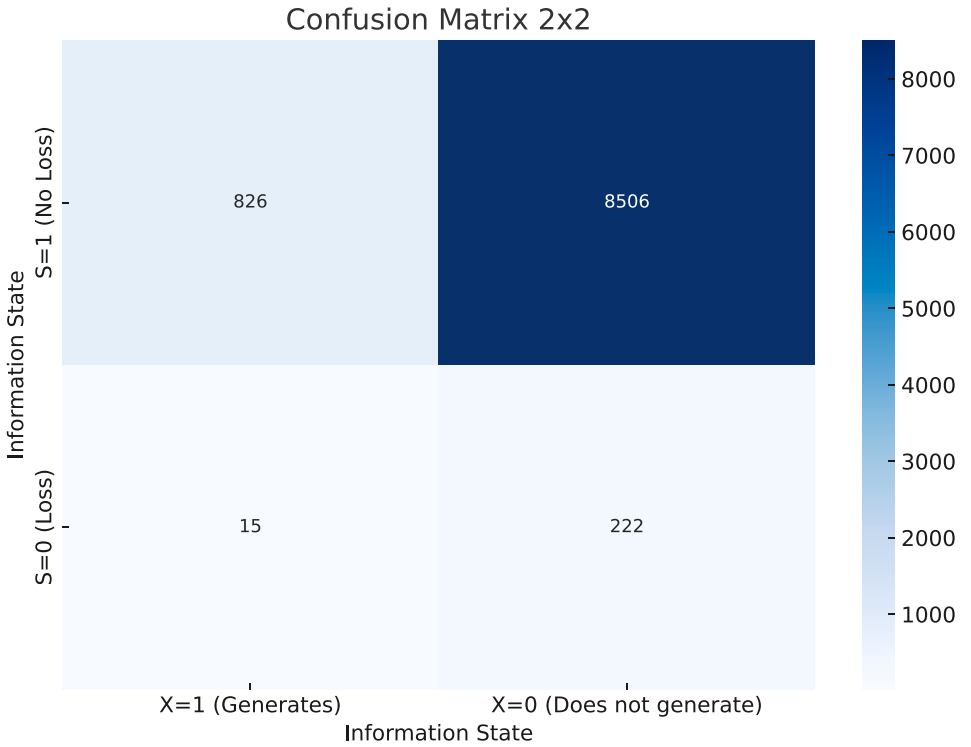

Fig. 11. Confusion matrix for the probabilistic knowledge-generation model under a synthetic scenario with 10,000 nodes. The matrix summarizes estimated absolute values for the relationship between research data availability (loss vs. no loss) and the generation of new knowledge, based on model assumptions rather than empirical observations. This representation is used as an interpretative tool to illustrate sensitivity, specificity, and error trade-offs implied by the probabilistic model, rather than as a validation against real-world data.

To quantify this effect, we define the *population cost*, which represents the expected reduction in knowledge production attributable to research data loss. The detailed derivation is presented in Appendix E, but in practical terms, for our dataset of 13,206 affected articles, the estimated loss corresponds to approximately 1072 publications. In other words, nearly one in every eight potentially productive works fails to generate new knowledge due to the unavailability of critical referenced research data Table G.1.

Fig. 11 visualizes this relationship through a confusion matrix that combines the states of information loss and knowledge generation. The distribution clearly illustrates the asymmetry between accessible and inaccessible conditions: when all references remain available, about 9.7% of articles contribute new knowledge, whereas in cases of information loss this share decreases to around 6.8%.

The formal derivation of the population cost and its numerical estimation are reported in Appendix E.

5. Analysis

5.1. Economic models analysis

The economic models applied in this study provide a quantitative interpretation of information loss that complements the network-based results. Scientific knowledge production can be viewed as a cumulative process in which datasets and referenced resources function as intermediate inputs; their unavailability, therefore, constitutes a negative shock to the knowledge production system.

Rather than estimating monetary losses, the models map citation-based impact patterns onto established economic concepts such as opportunity cost, productivity loss, and negative externalities. In this sense, they should be interpreted as structural proxies that translate network effects into economically meaningful indicators.

The opportunity cost formulation yields a conservative lower-bound estimate of forgone scientific output, capturing the loss of potential reuse associated with inaccessible data. The adapted Solow–Swan framework extends this view by linking information availability to aggregate research productivity, treating datasets as a form of intangible capital whose depreciation reduces the effective output of the research workforce. Complementarily, the population cost derived from sensitivity and specificity analysis quantifies the expected reduction in knowledge generation under uncertainty Fig. H.2.

Together, these models indicate that information loss affects scientific production through multiple channels: by limiting immediate reuse, reducing system-wide productivity, and increasing the likelihood that potentially productive research fails to materialize. Table 10 summarizes the estimated losses across models.

Table 10
Summary of estimated losses by economic model.

Model	Estimated Loss	Unit
Opportunity Cost	473.2	Scientific outputs
Population Cost (Sensitivity & Specificity)	1,072.0	Ungenerated articles

The economic estimations presented in this section build directly upon the structural propagation results obtained from the citation network analysis. In particular, the number of affected nodes identified through the propagation models is interpreted as a proxy for the potential scale of knowledge dependencies influenced by the loss of a given dataset.

Under this interpretation, each affected publication represents a potential unit of scientific production whose reproducibility, reuse, or extension may be compromised by the unavailability of the referenced resource. The economic models therefore translate the structural propagation of affected publications into estimates of opportunity cost in scientific production. Consequently, the economic estimates presented in this work should be interpreted as approximations based on structural proxies of knowledge dependency rather than as precise measurements of actual dataset reuse.

5.2. Structural analysis of the citation network

Network-based propagation models provide the structural foundation for these economic interpretations by revealing how localized information loss propagates through citation networks. Across models, propagation patterns exhibit strong non-linearities: a small number of structurally central articles act as superpropagators, generating disproportionate downstream effects when affected.

This behavior mirrors amplification mechanisms observed in economic production networks, where shocks to central inputs propagate widely across dependent units. At the same time, the observed decoupling between structural centrality and received impact highlights the existence of distinct roles within the scientific system. Some articles primarily absorb informational loss, acting as risk sinks, while others represent latent sources of systemic vulnerability whose current exposure underestimates their potential future impact.

These findings underscore the importance of network position in mediating the consequences of information loss and highlight the limitations of assessments based solely on local or article-level indicators.

5.2.1. Gozinto, SAIS, affectation, sensitivity & specificity, and transmission

The citation network extracted from unarXive exhibits a high degree of structural resilience: in most models, propagation remains limited beyond depth 3. Gozinto produces the most extensive propagation, while the Affectation–Transmission model constrains diffusion at shallow depths. The SAIS model yields intermediate behavior, with saturation effects emerging at depth 4.

Although all models originate from the same set of initially affected nodes, their propagation profiles differ substantially, reflecting contrasting assumptions about transmission, attenuation, and recovery. This diversity highlights the sensitivity of impact estimates to modeling choices and reinforces the value of comparative, multi-model analysis.

5.3. Policy relevant insights

Taken together, the economic and structural analyses indicate that the costs of information unavailability are unevenly distributed across the scientific system. Targeted interventions-such as prioritizing preservation efforts for datasets associated with structurally central articles-may therefore yield disproportionate benefits. From a policy perspective, these results support a shift from uniform preservation strategies toward risk-aware approaches informed by network structure and propagation potential.

6. Limitations

This study has several limitations that should be considered when interpreting the results.

First, the identification of articles relying on datasets is inherently challenging, as data references are not always explicitly labeled and may appear in the main text, footnotes, or supplementary materials. Constructing a fully exhaustive corpus would require large-scale text mining and semantic analysis beyond the scope of this work. Instead, we rely on a curated OpenAlex dataset, which represents a widely used and systematically collected source and is therefore appropriate for large-scale, comparative analysis. While this approach may miss some dataset usages, it provides a consistent and reproducible basis for studying structural propagation patterns Fig. H.3.

Second, the estimation of broken or inaccessible references may be conservative due to technical limitations in the automated link-checking process. These include misconfigured servers returning incorrect error codes, anomalous responses to automated agents (e.g., status code 418, "*I'm a teapot*"), and temporary ISP-level restrictions in Spain that blocked access to major CDNs, potentially flagging legitimate content as unavailable (Radauskas, 2024). Importantly, these effects are expected to introduce a systematic bias in absolute counts rather than altering the relative structure of the citation network. As a result, while the reported number of inaccessible references may be overestimated, the comparative propagation patterns and model-based rankings remain robust.

Future work will address this limitation by incorporating hybrid validation strategies that combine automated link checking with archival services (e.g., web archives) and targeted manual verification for critical nodes, improving both recall and precision in availability assessment.

A further limitation concerns the use of citation networks as structural proxies for knowledge dependencies. Citation relationships capture structural connections between publications but do not necessarily reveal the precise role of the cited resource. In particular, a citation to an article associated with a dataset does not guarantee that the dataset itself was directly used, reused, or provided by the citing work. Consequently, the propagation models employed in this study should be interpreted as identifying potential knowledge dependencies rather than confirmed instances of dataset reuse.

For this reason, our analysis focuses on structural propagation patterns rather than on the semantic role of individual citations. In this context, the FAIR Data Principles and research accessibility provide the conceptual motivation for the study: even when the precise function of a cited resource cannot be determined, the loss of access to referenced research objects may still undermine verification, reuse, and reproducibility within the scientific record.

Future research could address this limitation by integrating semantic analysis of scientific articles, such as citation context analysis and natural language processing techniques applied to full-text corpora. Recent advances in large-scale text mining, transformer-based language models, and information extraction methods make it increasingly feasible to identify the functional role of citations (e.g., dataset usage, methodological reference, or background citation). Combining such semantic signals with structural citation networks could enable more precise identification of dependency relations between scientific outputs and the research resources on which they rely.

A potential concern when using arXiv data is that preprints may vary in terms of editorial quality and reference curation. In particular, lower-quality submissions could potentially exhibit higher rates of link failures due to less rigorous referencing practices. However, the focus of this study is on structural propagation mechanisms in citation networks rather than on the epistemic quality of individual publications. Moreover, the consistency of the results observed in the complementary OpenAlex based analysis suggests that the main propagation patterns identified here are not driven by characteristics specific to arXiv preprints Fig. .

The use of arXiv in this study is primarily motivated by methodological advantages such as full text availability and explicit URL annotations. To mitigate potential biases associated with preprint literature, we complement the arXiv analysis with citation networks reconstructed from OpenAlex, which aggregates peer reviewed publications across multiple publishers. The consistency of structural propagation patterns observed across both datasets suggests that the main conclusions are not driven by specific characteristics of arXiv preprints.

Finally, the modeling of research data loss propagation in scientific citation networks remains an emerging research area. While the models applied here capture key structural and economic mechanisms, further methodological development is needed, particularly to integrate empirical validation, author-level effects, and domain-specific citation practices. The present study should therefore be understood as an initial step toward a more comprehensive quantitative framework for assessing the systemic consequences of reproducibility failures Fig. H.1.

7. Conclusions

This study examined the systemic consequences of dataset unavailability in scientific citation networks through a combination of economic, probabilistic, and network-based models. By treating citations as observable proxies for knowledge reuse and dependency, we analyzed how the loss of access to research resources propagates through interconnected bodies of scientific work.

Our results show that information loss is not merely a local reproducibility issue but a structural phenomenon with system-wide implications. Network-based analyses reveal that the effects of dataset unavailability depend strongly on the position of affected articles within the citation network, giving rise to non-linear propagation patterns and the emergence of superpropagators. These findings echo insights from economic theories of production networks, where shocks to central inputs generate disproportionate downstream losses.

Economic models adapted in this work translate these structural effects into interpretable cost proxies. Opportunity cost estimates capture foregone reuse, productivity-based formulations link information loss to reduced research capacity, and probabilistic models quantify the expected decline in knowledge generation under uncertainty. While these models are intentionally stylized and not calibrated for prediction, their convergence highlights a robust qualitative conclusion: the unavailability of research data imposes a measurable efficiency loss on the scientific system.

Importantly, the combination of economic interpretation and network structure suggests that mitigation strategies should be selective rather than uniform. Preserving or restoring access to datasets associated with structurally central publications may prevent cascading losses that far exceed the local cost of data recovery. This insight provides a quantitative rationale for prioritizing preservation efforts based on network position and propagation risk.

Several extensions could refine this framework. Incorporating empirically calibrated attenuation mechanisms, distinguishing functional roles of citations, and integrating author- or institution-level dynamics would allow for closer alignment with economic models of innovation and knowledge diffusion. Nonetheless, even in its current form, the approach demonstrates how economic concepts can be meaningfully operationalized using citation network structure.

Overall, this work contributes to the literature on the economics of science by providing a quantitative framework to assess how failures in data accessibility translate into systemic inefficiencies. It reinforces the view that sustained scientific progress depends not only on producing new knowledge, but also on preserving the informational infrastructure that enables cumulative discovery.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used OpenAI ChatGPT, 2024 version and Google Gemini in order to improve language clarity, grammar and spelling, and to assist in the visual standardization of tables. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

CRediT authorship contribution statement

Jorge Chamorro-Padial: Writing – review & editing, Writing – original draft, Software, Resources, Methodology, Investigation, Data curation, Conceptualization; **Francisco-Javier Rodrigo-Ginés:** Writing – review & editing, Writing – original draft, Funding acquisition, Formal analysis, Data curation; **Rosa Rodríguez-Sánchez:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Conceptualization; **R. M. Gil:** Writing – review & editing, Writing – original draft, Validation, Project administration, Methodology, Investigation; **Roberto García:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Funding acquisition, Formal analysis.

Data availability

The resulting dataset is published at Chamorro Padial et al. (2025a). The Python code created to process the data of this article is published at Chamorro Padial et al. (2025b).

Acknowledgment

This work was partially supported by the project “ANGRU: Applying kNowledge Graphs to research data ReUsability” with reference PID2020-117912RB-C22 and funded by MCIN/AEI/10.13039/501100011033.

Appendix A. Gozinto theorem: Toy example

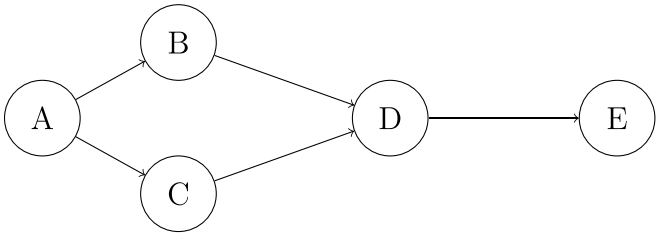

Fig. A.1. Toy citation graph. Node A is affected.

Table A.1
Gozinto Toy example: Total received impact for each node.

Node	Impact received
A	1.0
B	0.5
C	0.5
D	0.5
E	0.25

The impact received by each node is summarized in Table A.1. Fig. A.1 illustrates how the Gozinto model captures not only direct citations from affected nodes, but also indirect and cumulative influence across multiple layers.

We assume that node A is affected and starts with an impact of 1. This impact propagates through the network with a decay factor of $\alpha = 0.5$ per level of depth. The propagation is illustrated up to a maximum depth of 3 for clarity and interpretability. This choice is purely illustrative and does not imply convergence of the propagation process, but rather highlights how impact accumulates across successive citation generations. The propagation unfolds as follows:

- Level 1: A → B and A → C each receive $1 \times 0.5 = 0.5$
- Level 2:
  - B → D: $0.5 \times 0.5 = 0.25$
  - C → D: $0.5 \times 0.5 = 0.25$
  - ⇒ D receives $0.25 + 0.25 = 0.5$
- Level 3: D → E: $0.5 \times 0.5 = 0.25$

Appendix B. Mathematical formulation of the FPSC model

The Forward Path Search Count (FPSC) metric can be expressed as:

$$FPSC(p_i; \phi, K) = \sum_{\forall (p_j, p_i), k=1}^{K} N^{-}(p_j, p_i, k) \cdot \phi(k)$$

where:

- $N^{-}(p_j, p_i, k)$ represents the number of influence paths of length k passing through article p_i from p_j;
- $\phi(k)$ is the decay factor applied to paths of length k;
- K defines the maximum citation path length considered.

Simplified variants of the model omit the decay term and/or the maximum path length constraint.

Appendix C. Mathematical formulation of the SAIS model

The SAIS model represents the transitions between the states of Susceptible (S), Alert (A), and Infected (I) articles through the following system of differential equations:

$$\frac{dS}{dt} = -\beta S I + \delta A$$

$$\frac{dA}{dt} = \beta S I - \kappa A - \delta A$$

$$\frac{dI}{dt} = \kappa A$$

where:

- β is the transmission rate from infected to susceptible articles;
- κ is the infection rate from the alert state;
- δ is the rate at which alert articles return to the susceptible state.

An optional recovery parameter γ may be included to model the probability that an infected article becomes accessible again. In the present study we assume $\gamma = 0$, implying no recovery of lost references.

Appendix D. Affectation and transmission model

We represent an article P as a set of sections:

$$P = \{s_i \mid i \in \{1, 2, \ldots, n\}\},$$

where s_i denotes the ith section and n is the total number of sections.

Each article contains a set of references R, and a binary relation $\mathcal{R}$ links sections to the references they contain:

$$\mathcal{R} = \{(s_i, r_j) \mid s_i \in S, r_j \in R, r_j \in R_i\}.$$

The subset of references within a given section is:

$$R_i = \{r_j \in R \mid (s_i, r_j) \in \mathcal{R}\}.$$

Each section is assigned a weight $w_i \in [0, 1]$ through a function:

$$w : S \to W, \qquad \sum_{i=1}^{n} w_i = 1.$$

Following (Klein et al., 2014), we define the possible states of a reference as $E = \{\text{“Health”}, \text{“Infected”}\}$, and a mapping:

$$e : R \to E.$$

A section is considered infected if at least one of its references is infected:

$$s_i \text{ is infected } \Leftrightarrow \bigvee_{r_j \in R} \big((s_i, r_j) \in \mathcal{R} \wedge e(r_j) = \text{“Infected”}\big).$$

The affectation level of an article A is given by the weighted sum of infected sections:

$$A = \sum_{s_i \in S} w_i \cdot I(s_i), \quad I(s_i) = \begin{cases} 1 & \text{if } s_i \text{ is infected,} \\ 0 & \text{otherwise.} \end{cases}$$

Hence, $A \in [0, 1]$ represents the proportion of the article that is compromised.

Extension: Transmission across a Citation Network

To capture propagation between articles, let $C = \{p_1, p_2, \ldots, p_n\}$ denote the corpus of articles and L^{ext} the set of external references. The binary affectation function for a reference r is defined as:

$$A_{binary}(r) = \begin{cases} 0, & \text{if } e(r) = \text{“Health”,} \\ 1, & \text{if } e(r) = \text{“Infected”.} \end{cases}$$

For each section s_i, we define a transmission indicator $I_T(s_i)$ that combines direct external infections and propagated ones through the Linear Threshold Model (LTM):

$$I_T(s_i) = \max_{r_i \in L^{ext}_{s_i}} A_{binary}(r_i) \vee LTM(s_i),$$

where $L^{ext}_{s_i} \subseteq L^{ext}$ and:

$$LTM(s_i) = \begin{cases} 1, & \text{if } \sum_{r_j \in C_{s_i}} w^t_{r_j} A_t(r_j) \geq \theta, \\ 0, & \text{otherwise.} \end{cases}$$

Here,

- A_t is the transmitted affectation value,
- $w^t_{r_j}$ is the weight of reference r_j, defined as $w^t_{r_j} = \frac{1}{|R_{s_i}| + \epsilon}$,
- θ is the global threshold defining activation,
- $C_{s_i} \subseteq C$ are the internal corpus references within section s_i,
- ϵ is a small constant to prevent division by zero.

The overall transmission-based affectation for an article p is then:

$$A_T(p) = \sum_{s_i \in S_p} w(s_i) \cdot I_T(s_i),$$

with a recursive stopping condition when a cited article is not part of the corpus C.

Appendix E. Formal definition of the specificity and sensitivity model

Let $X \in \{0, 1\}$ denote the binary outcome of knowledge generation, where:

$$X = \begin{cases} 1, & \text{if new knowledge is generated,} \\ 0, & \text{otherwise.} \end{cases}$$

Let $S \in \{0, 1\}$ represent the state of information integrity:

$$S = \begin{cases} 1, & \text{if information is preserved (no loss),} \\ 0, & \text{if information is lost or inaccessible.} \end{cases}$$

We define:

- **Sensitivity** (S_1) as the conditional probability of generating new knowledge given that information is intact:

$$S_1 = P(X = 1 \mid S = 1).$$

- **Specificity** (S_0) as the conditional probability of not generating new knowledge given that information has been lost:

$$S_0 = P(X = 0 \mid S = 0).$$

Applying Bayes' theorem, sensitivity can be expressed as:

$$P(X = 1 \mid S = 1) = \frac{P(S = 1 \mid X = 1)\, P(X = 1)}{P(X = 1)P(S = 1 \mid X = 1) + P(X = 0)P(S = 1 \mid X = 0)}.$$

Similarly, specificity is given by:

$$P(X = 0 \mid S = 0) = \frac{P(X = 0)P(S = 0 \mid X = 0)}{P(X = 1)P(S = 0 \mid X = 1) + P(X = 0)P(S = 0 \mid X = 0)}.$$

In this adapted framework, information loss occurs when one or more references in an article become inaccessible. Thus, by evaluating the reference set of a publication, both sensitivity and specificity can be estimated empirically. These measures describe how effectively new knowledge can be generated or hindered under varying conditions of data availability.

Estimation of Population Cost

Beyond the individual probabilities estimated by the *Specificity and Sensitivity Model*, we define the *population cost C* as the expected reduction in knowledge production caused by information loss across the citation network. This measure captures the aggregate effect of inaccessible information on the system's overall ability to generate new knowledge.

Formally, the population cost is defined as:

$$C = \left(P(X_1 \mid S_1) - P(X_1 \mid S_0)\right) \cdot N_{S_0},$$

where:

- $P(X_1 \mid S_1)$ is the probability of generating new knowledge when all references are available;
- $P(X_1 \mid S_0)$ is the probability of generating new knowledge when information loss occurs;
- N_{S_0} is the number of nodes (articles) affected by information loss.

The term $\Delta P = P(X_1 \mid S_1) - P(X_1 \mid S_0)$ represents the marginal reduction in knowledge generation due to the transition from complete to incomplete information states. The resulting product with N_{S_0} provides an estimate of the total number of articles whose capacity to generate new knowledge is compromised.

Applying this formulation to the empirical data yields:

$$P(X_1 \mid S_1) = 0.0826, \qquad P(X_1 \mid S_0) = 0.0015,$$

so that

$$\Delta P = 0.0826 - 0.0015 = 0.0811.$$

Given $N_{S_0} = 13{,}206$, the estimated population cost is:

$$C = 0.0811 \times 13{,}206 \approx 1{,}072.$$

Hence, approximately 1072 articles in the analyzed dataset may have failed to generate new knowledge due to the loss of critical referenced information. This value provides a first-order estimate of the systemic impact of data inaccessibility on scientific productivity within the studied network.

Appendix F. Datasets temporarily or permanently unavailable

Table F.2 summarises the datasets that were identified as temporarily or permanently unavailable during the manual verification process. For each dataset, the corresponding article citing the resource, the observed access status, and the original publication context are reported.

The accessibility of these resources was manually checked on three separate occasions (13 March, 12 April, and 14 May 2023) in order to reduce the likelihood that temporary server issues or short-term outages would lead to false positives. A dataset was considered unavailable when the referenced resource could not be accessed or retrieved through the URI provided in the original publication.

This list provides empirical evidence of the types of external resources that may become inaccessible over time, including project webpages, personal webpages, and repository-hosted materials. Such cases illustrate how the disappearance or temporary unavailability of referenced datasets may contribute to the loss of reproducibility and the propagation of incomplete knowledge within citation networks.

Table F.2
List of datasets temporarily or permanently unavailable. Access checked three times (13 March, 12 April, 14 May 2023).

Dataset name or URI	Paper	Comment	Published on
Queensland Suicide Register	(Arnautovska et al., 2015)	Temporarily inaccessible	Project webpage
National Coroners Information System	(Arnautovska et al., 2015)	Temporarily inaccessible	Project webpage
https://traps.ncipmc.org	(Rosenheim & Gratton, 2017)		Project webpage
Soybean dataset	(Rosenheim & Gratton, 2017)		Project webpage
European Pollen database	(Fyfe et al., 2015)		Project webpage
https://www.naturalearthdta.com/	(Bleasdale et al., 2020)		Project webpage
https://portal.mtt.fi/portal/page/portal/kasper/pelto/peltopalvelut/lajikekokeet	(Peltonen-Sainio et al., 2016)		Personal homepage
https://nadp.slh.wisc.edu/MDN/maps.aspx	(Yu et al., 2019)		Project webpage
https://www.pbl.nl/en/publications/2006/N2OAndNOEmissionFrom	(Philibert et al., 2012)		Project webpage
https://github.com/liuliu66/AgriPest	(Wu & Ma, 2022)		GitHub
Sen4Agrinet	(D. Sykas et al., 2022)		Project webpage
US county-level agricultural crop production typology	(Wagner et al., 2019)	Temporarily removed	FigShare

Appendix G. Gozinto arXiv ids and their corresponding references

Table G.1
arXiv ids and their corresponding papers.

arXiv id	Title	Reference
2102.10772	UniT: Multimodal Multitask Learning with a Unified Transformer	(Hu & Singh, 2021)
2102.10274	Concealed Object Detection	(Fan et al., 2022)
2212.08059	Rethinking Vision Transformers for MobileNet Size and Speed	(Li et al., 2023)
2208.07658	DRAGON: Decentralized Fault Tolerance in Edge Federations	(Tuli et al., 2022b)
2207.05501	Next-ViT: Next Generation Vision Transformer for Efficient Deployment in Realistic Industrial Scenarios	(Li et al., 2022)
2203.07140	CAROL: Confidence-Aware Resilience Model for Edge Federations	(Tuli et al., 2022a)
2111.13923	Learning A 3D-CNN and Transformer Prior for Hyperspectral Image Super-Resolution	(Ma et al., 2021)

Appendix H. Gozinto results

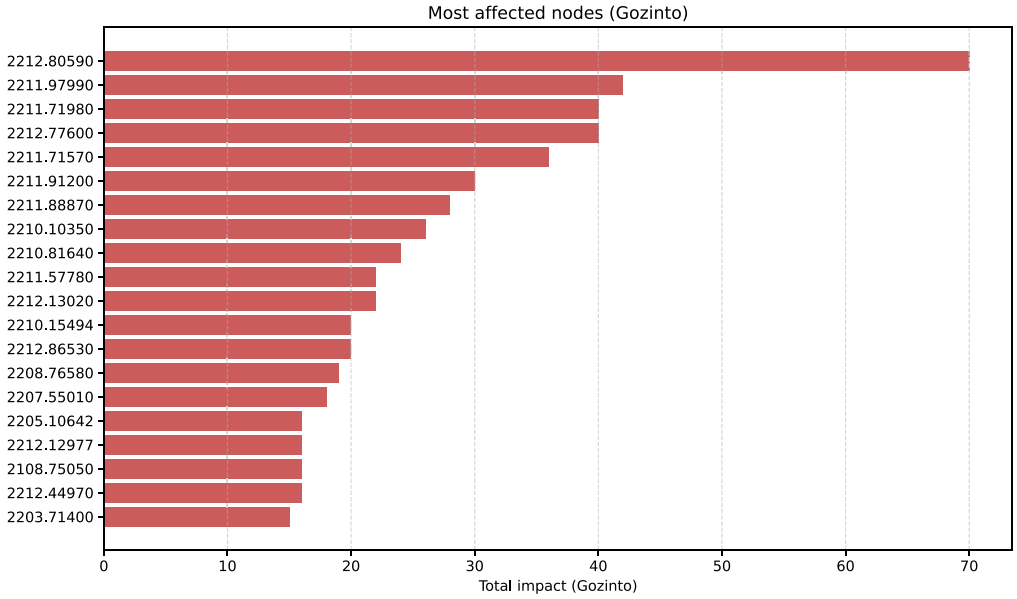

Fig. H.1. Nodes receiving the highest cumulative impact according to the Gozinto model. Bars represent the total impact received from affected citation paths, highlighting articles that are structurally most exposed to upstream dataset unavailability. Comparison with propagated-impact rankings shows that highly affected nodes do not necessarily act as superpropagators, underscoring the presence of distinct structural roles within the citation network.

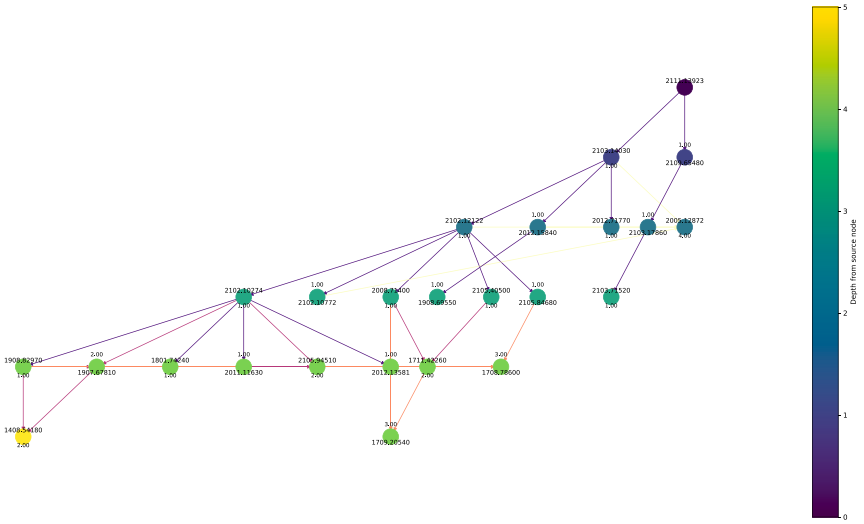

Fig. H.2. Citation subnetwork induced by the superpropagator .13923. Nodes are arranged according to citation depth along directed paths originating from the superpropagator. Node color encodes the cumulative propagated impact, illustrating how influence is transmitted across multiple citation branches. The structure highlights the role of superpropagators as central conduits through which dataset unavailability can generate widespread downstream effects.

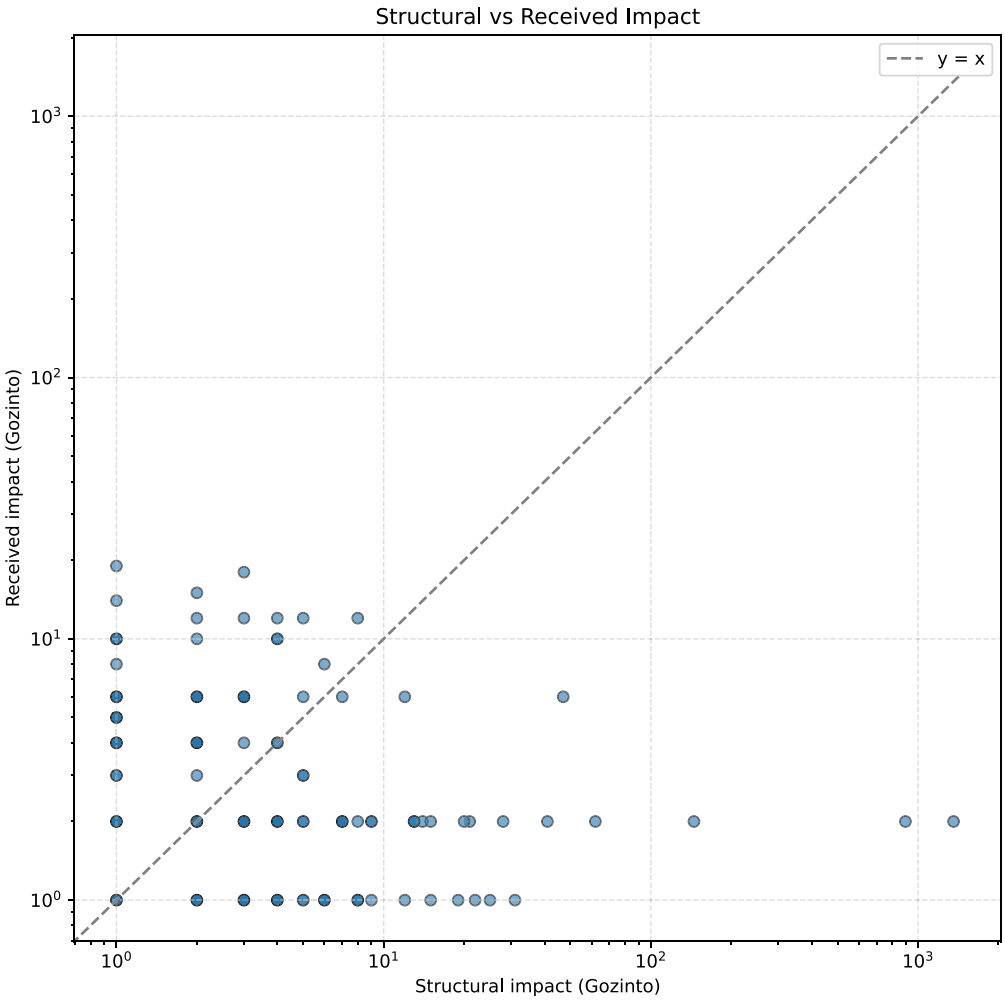

Fig. H.3. Relationship between structural (propagated) impact and cumulative received impact for citation nodes, based on the Gozinto model. Both axes are shown on a logarithmic scale. The diagonal indicates the hypothetical case where received and propagated impact are equal. The wide dispersion of points around this line reveals a clear decoupling between vulnerability and structural influence: nodes that receive high impact are not necessarily those that propagate it further, and vice versa. This pattern highlights the presence of distinct structural roles within the citation network. Such decoupling supports the use of multiple complementary impact proxies in the analysis.

Appendix I. Sensitivity analysis with respect to traversal depth

To assess whether the qualitative conclusions of the propagation models depend on the chosen traversal depth, we extended the exploration depth to values up to 6 in selected OpenAlex case studies. The results show that while the absolute number of affected nodes increases with depth, the overall propagation patterns remain stable and we reach a saturation point. In particular, the ranking of the most influential nodes and the skewed distribution of propagated impact remain consistent.

As shown in Table I.2, the number of affected nodes grows rapidly when increasing the traversal depth from 3 to 5. Extending the depth from 5 to 6 still increases the number of affected nodes in some cases. However, these additional nodes mostly correspond to more distant parts of the citation network and do not significantly alter the qualitative propagation patterns observed in the main analysis.

Table I.2
Sensitivity analysis of the propagation depth parameter for a representative set of OpenAlex case studies using the Gozinto propagation model. The results show that the number of affected nodes generally increases with traversal depth, and although the rate of growth tends to decrease at higher depths, substantial quantitative differences may still appear between depth 5 and depth 6 in some cases.

OpenAlex ID	Depth 5	Depth 6
W2031816812	481,069	1,106,458
W2336803818	53,647	187,764
W2549187017	0	0
W4225728415	0	0
W3097366679	4076	13,450
W2970491196	0	0

Appendix J. Manual verification of OpenAlex case studies

To validate the automated identification of inaccessible research resources, we conducted a manual verification of the six case studies selected from the OpenAlex citation network analysis. These case studies correspond to articles identified as referencing datasets or data infrastructures whose accessibility status was detected through the automated link-checking procedure.

For each article, we manually inspected the full text to identify the referenced dataset, data portal, or external research resource associated with the detected link. The original URL cited in the publication was then accessed manually (March 2026) in order to verify whether the resource remained accessible, had migrated to a different location, or was no longer available.

This manual verification allowed us to confirm the accuracy of the automated detection process and to identify the specific causes of inaccessibility. The observed issues included discontinued databases, migration of data infrastructures, removal of project webpages, and typographical errors in cited URLs. Importantly, these cases illustrate the different mechanisms through which references to research resources may become inaccessible over time.

Table J.1 summarizes the six manually verified case studies, including the article identifier, the referenced dataset or resource, the original cited URL, and its accessibility status at the time of verification.

Table J.1
Manual verification of the six OpenAlex case studies used to validate the automated detection of inaccessible research resources. For each article, the referenced dataset or research infrastructure was identified and the accessibility of the original cited URI was verified manually. All URLs were accessed and verified on March 7, 2026.

Case	OpenAlex ID	Article title	Referenced resource	Original URI	Status
1	W2970491196	US county-level agricultural crop production typology	County-level crop production typology dataset (Figshare)	https://figshare.com/articles/dataset/US_county-level_agricultural_crop_production_typology/8132867/2	Accessible
2	W2031816812	Quantifying Uncertainties in N_2O Emission Due to N Fertilizer Application in Cultivated Areas	Stehfest and Bouwman emission dataset (PBL Netherlands Environmental Assessment Agency)	http://www.pbl.nl/en/publications/2006/N2OAndNOEmission-FromAgriculturalFieldsAndSoil-sUnderNaturalVegetation	Inaccessible
3	W2336803818	Ecoinformatics (Big Data) for Agricultural Entomology: Pitfalls, Progress, and Promise	VegBank vegetation database	http://vegbank.org	Inaccessible
4	W2549187017	Land Use, Yield and Quality Changes of Minor Field Crops: Is There Superseded Potential to Be Reinvented in Northern Europe?	FAOSTAT crop production database portal	http://faostat3.fao.org/browse/Q/QC/E	Inaccessible
5	W4225728415	A Sentinel-2 Multiyear, Multicountry Benchmark Dataset for Crop Classification and Segmentation With Deep Learning	Copernicus Open Access Hub API (Sentinel-2 data access infrastructure)	https://scihub.copernicus.eu/twiki/do/view/SciHubWebPortal/APIHub	Inaccessible
6	W3097366679	Isotopic and microbotanical insights into Iron Age agricultural reliance in the Central African rainforest	Natural Earth geographic dataset	https://www.naturalearthdta.com/downloads/	Inaccessible

Note: All URLs were accessed and verified manually on March 7, 2026.

Table J.2
Propagation estimates for the manually reconstructed citation layers of case study W2031816812. The Affectation–Transmission model follows the structural citation expansion at early depths, while the SAIS model produces a more constrained propagation consistent with the ratios observed in the main OpenAlex experiments.

Depth	Citation count	Gozinto	Affectation–Transmission	SAIS
1	89	89	89	45
2	7051	7051	7051	3342

Table J.3
Propagation estimates for the manually reconstructed citation layers of case study W2336803818. The Affectation–Transmission model follows the structural citation expansion at early depths, while the SAIS model produces a more constrained propagation consistent with the ratios observed in the main OpenAlex experiments.

Depth	Citation count	Gozinto	Affectation–Transmission	SAIS
1	144	144	144	72
2	14,052	14,052	14,052	6662

Table J.4
Propagation estimates for the manually reconstructed citation layers of case study W2336803818. The Affectation–Transmission model follows the structural citation expansion at early depths, while the SAIS model produces a more constrained propagation consistent with the ratios observed in the main OpenAlex experiments.

Depth	Citation count	Gozinto	Affectation–Transmission	SAIS
1	39	39	39	20
2	4014	4014	4014	1903

Table J.5
Propagation estimates for the manually reconstructed citation layers of case study W4225728415. The Affectation–Transmission model follows the structural citation expansion at early depths, while the SAIS model produces a more constrained propagation consistent with the ratios observed in the main OpenAlex experiments.

Depth	Citation count	Gozinto	Affectation–Transmission	SAIS
1	61	61	61	31
2	6532	6532	6532	3097

Table J.6
Propagation estimates for the manually reconstructed citation layers of case study W3097366679. The Affectation–Transmission model follows the structural citation expansion at early depths, while the SAIS model produces a more constrained propagation consistent with the ratios observed in the main OpenAlex experiments.

Depth	Citation count	Gozinto	Affectation–Transmission	SAIS
1	87	87	87	44
2	7383	7383	7383	3500

To complement the manual verification of dataset accessibility described in this appendix, we estimated propagation metrics for several representative OpenAlex case studies corresponding to the manually verified dataset references.

For each case, citation layers were manually reconstructed up to depth 2 using the OpenAlex API, and the resulting citation counts were manually inspected and recorded. Based on these reconstructed citation layers, we estimated the corresponding propagation values for the Gozinto, Affectation–Transmission, and SAIS models using the same propagation assumptions applied in the OpenAlex experiments presented in the main text.

The resulting estimates are reported in Tables J.2–J.6. As expected, the Gozinto and Affectation–Transmission models follow the structural expansion of the citation network at early depths, whereas the SAIS model produces a more constrained propagation pattern consistent with the behaviour observed in the large-scale experiments.

Because the citation layers were reconstructed manually, the reported values should be interpreted as approximate estimates derived from the observed citation expansion rather than exact outputs of the full network models. Their purpose is to provide a structural consistency check for the manually verified case studies.

References

Abby, M. (1994). Peer review is an effective screening process to evaluate medical manuscripts. *JAMA: The Journal of the American Medical Association*, *272*(2), 105. https://doi.org/10.1001/jama.1994.03520020031008

Antelman, K. (2004). Do open-access articles have a greater research impact? *College Research Libraries*, *65*(5), 372–382. https://doi.org/10.5860/crl.65.5.372

Arnautovska, U., McPhedran, S., & De Leo, D. (2015). Differences in characteristics between suicide cases of farm managers compared to those of farm labourers in Queensland, Australia. *Rural and Remote Health*, *15*(3). QID: Q38590437. https://doi.org/10.22605/RRH3250

Ashley, K. (2019). Using open and FAIR data to increase research efficiency. SPARC Europe. https://sparceurope.org/using-open-and-fair-data-to-increase-research-efficiency/. Report. Accessed: 2025-05-07. Archived at: https://web.archive.org/web/20260423215604/https://sparceurope.org/download/7440/?tmstv=1776981345.

Bleasdale, M., Wotzka, H.-P., Eichhorn, B., Mercader, J., Styring, A., Zech, J., Soto, M., Inwood, J., Clarke, S., Marzo, S., Fiedler, B., Linseele, V., Boivin, N., & Roberts, P., et al. (2020). Isotopic and microbotanical insights into iron age agricultural reliance in the central African rainforest. *Communications Biology*, *3*(1). QID: Q101038811. https://doi.org/10.1038/s42003-020-01324-2

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. https://doi.org/10.1088/1742-5468/2008/10/P10008

Buchanan, J. M. (2018). Opportunity cost. In *The new palgrave dictionary of economics* , pp. 9822–9826). London: Palgrave Macmillan UK. https://doi.org/10.1057/978-1-349-95189-5_1433

Carnevale, R., & Aronsky, D. (2007). The life and death of URLs in five biomedical informatics journals. *International Journal of Medical Informatics*, *76*(4), 269–273. https://doi.org/10.1016/j.ijmedinf.2005.12.001

Chamorro-Padial, J., García, R., & Gil, R. (2024). A systematic review of open data in agriculture. *Computers and Electronics in Agriculture*, *219*, 108775. https://doi.org/10.1016/j.compag.2024.108775

Chamorro Padial, J. Rodrigo-Ginés, F.-J. Rodríguez Sánchez, R. M. Gil Iranzo, R. M. García González, R. (2025). Replication Data for: Modeling the impact of research data unavailability on science. CORA.Repositori de Dades de Recerca. https://doi.org/10.34810/data2382.

Chamorro Padial, J. Rodrigo-Ginés, F.-J. Rodríguez Sánchez, R. M. Gil Iranzo, R. M. García González, R. (2025). Scripts for: Modeling the impact of research data unavailability on science. CORA.Repositori de Dades de Recerca. https://doi.org/10.34810/data2383.

Costa, M., Gomes, D., & Silva, M. J. (2017). The evolution of web archiving. *International Journal on Digital Libraries*, *18*(3), 191–205. https://doi.org/10.1007/s00799-016-0171-9

SykasD., SdrakaM., ZografakisD., & PapoutsisI. (2022). A Sentinel-2 multiyear, multicountry benchmark dataset for crop classification and segmentation with deep learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *15*, 3323–3339. QID: Q114085977. https://doi.org/10.1109/JSTARS.2022.3164771

De Martino, G., & Spina, S. (2015). Exploiting the time-dynamics of news diffusion on the internet through a generalized susceptible-infected model. *Physica A: Statistical Mechanics and its Applications*, *438*, 634–644. https://doi.org/10.1016/j.physa.2015.07.022

Dellavalle, R. P., Hester, E. J., Heilig, L. F., Drake, A. L., Kuntzman, J. W., Graber, M., & Schilling, L. M. (2003). Going, going, gone: Lost internet references. *Science*, *302*(5646), 787–788. https://doi.org/10.1126/science.1088234

Dellinger, A. B. (2005). Validity and the review of literature. *Research in the Schools*, *12*(2), 41–54.

Di Cosmo, R. (2022). Should we preserve the world's software history, and can we? In G. Silvello, O. Corcho, P. Manghi, G. M. Di Nunzio, K. Golub, N. Ferro, & A. Poggi (Eds.), *Linking theory and practice of digital libraries* (pp. 3–7). Cham: Springer International Publishing (*vol. 13541*). Series Title: Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-031-16802-4_1

Ducut, E., Liu, F., & Fontelo, P., et al. (2008). An update on uniform resource locator (URL) decay in MEDLINE abstracts and measures for its mitigation. *BMC Medical Informatics and Decision Making*, *8*(1), 23. https://doi.org/10.1186/1472-6947-8-23

Elrashdi, A. S., & Elferjani, K. (2024). The potential of using IMaRD format to improve the effectiveness of scientific papers. *Int. J. Electr. Eng. and Sustain.*, *2*(3), 60–65. https://ijees.org/index.php/ijees/article/view/96.

Escamilla, E., Klein, M., Cooper, T., Rampin, V., Weigle, M. C., & Nelson, M. L., et al. (2022). The rise of GitHub in scholarly publications. In G. Silvello, O. Corcho, P. Manghi, G. M. Di Nunzio, K. Golub, N. Ferro, & A. Poggi (Eds.), *Linking theory and practice of digital libraries* (pp. 187–200). Cham: Springer International Publishing (*vol. 13541*). Series Title: Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-031-16802-4_15

Falagas, M. E., Karveli, E. A., & Tritsaroli, V. I. (2008). The risk of using the internet as reference resource: A comparative study. *International Journal of Medical Informatics*, *77*(4), 280–286. https://doi.org/10.1016/j.ijmedinf.2007.07.001

Fan, D.-P., Ji, G.-P., Cheng, M.-M., & Shao, L. (2022). Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(10), 6024-6042. https://doi.org/10.1109/tpami.2021.3085766

Funk, S., Gilad, E., Watkins, C., & Jansen, V. A. A., et al. (2009). The spread of awareness and its impact on epidemic outbreaks. *Proceedings of the National Academy of Sciences*, *106*(16), 6872–6877. https://doi.org/10.1073/pnas.0810762106

Fyfe, R. M., Woodbridge, J., & Roberts, N. (2015). From forest to farmland: Pollen-inferred land cover change across Europe using the pseudobiomization approach. *Global Change Biology*, *21*(3), 1197–1212. QID: Q46821518. https://doi.org/10.1111/gcb.12776

Garfield, E. (1964). "Science citation index"-a new dimension in indexing: This unique approach underlies versatile bibliographic systems for communicating and evaluating information. *Science*, *144*(3619), 649–654. https://doi.org/10.1126/science.144.3619.649

Ho, J. (2006). Hyperlink obsolescence in scholarly online journals. *Journal of Computer-Mediated Communication*, *10*(3), 00–00. https://doi.org/10.1111/j.1083-6101.2005.tb00263.x

Hu, R., & Singh, A. (2021). Unit: Multimodal multitask learning with a unified transformer. arXiv:2102.10772

Huang, S., Huang, Y., Bu, Y., Lu, W., Qian, J., & Wang, D. (2022). Fine-grained citation count prediction via a transformer-based model with among-attention mechanism. *Information Processing & Management*, *59*(2), 102799. https://doi.org/10.1016/j.ipm.2021.102799

Jiang, X., & Zhuge, H. (2019). Forward search path count as an alternative indirect citation impact indicator. *Journal of Informetrics*, *13*(4), 100977. https://doi.org/10.1016/j.joi.2019.100977

Johns Hopkins Center for Government Excellence (2016). Open data: How much does it cost? Johns Hopkins University. https://bloombergcities.jhu.edu/sites/default/files/bch-media/files/resources/civic_impact_open-data-how-much-does-it-cost.pdf. Archived at: https://web.archive.org/web/20240217170406/; https://bloombergcities.jhu.edu/sites/default/files/bch-media/files/resources/civic_impact_open-data-how-much-does-it-cost.pdf.

JOHN, H. C., Simisaye, A. O., & Iseyemi, T. J. (2024). Missing and recovery of URLs using internet archive: A case study on african journal of library archives and information science (AJLAIS). *International Journal of Information Science and Management (IJISM)*, (Online First). https://doi.org/10.22034/ijism.2024.1977641.0

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, *14*(1), 10–25. https://doi.org/10.1002/asi.5090140103

Klein, M., Van De Sompel, H., Sanderson, R., Shankar, H., Balakireva, L., Zhou, K., & Tobin, R., et al. (2014). Scholarly context not found: One in five articles suffers from reference rot. *PloS one*, *9*(12), e115253. https://doi.org/10.1371/journal.pone.0115253

Krauskopf, E., & Salgado, M. (2023). Inconsistency in the registration of the digital object identifier (DOI) of articles on web of science and scopus. *Investigación Bibliotecológica: Archivonomía, Bibliotecología e Información*, *37*(96), 129–144. https://doi.org/10.22201/iibi.24488321xe.2023.96.58784

Kulikov, D. A. (2019). The generalized solow model. In *Journal of physics: Conference series* (p. 012033). IOP Publishing (*vol. 1205*).

Lambiotte, R., Delvenne, J.-C., & Barahona, M. (2014). Random walks, markov processes and the multiscale modular organization of complex networks. *IEEE Transactions on Network Science and Engineering*, *1*(2), 76-90. https://doi.org/10.1109/tnse.2015.2391998
Lawrence, S., Pennock, D. M., Flake, G. W., Krovetz, R., Coetzee, F. M., Glover, E., Nielsen, F. A., Kruger, A., & Giles, C. L. (2001). Persistence of web references in scientific research. *Computer*, *34*(3), 26–31. https://doi.org/10.1109/2.901164
Lendvai, G. F., & Sasvári, P. (2025). 'Wasted' research and lost citations: A scientometric assessment of retracted documents in Scopus between 2001 and 2024. *Journal of Information Science*, (p. 01655515251362383). https://doi.org/10.1177/01655515251362383
Li, J., Xia, X., Li, W., Li, H., Wang, X., Xiao, X., Wang, R., Zheng, M., & Pan, X. (2022). Next-ViT: Next generation vision transformer for efficient deployment in realistic industrial scenarios. arXiv:2207.05501
Li, Y., Hu, J., Wen, Y., Evangelidis, G., Salahi, K., Wang, Y., Tulyakov, S., & Ren, J. (2023). Rethinking vision transformers for mobilenet size and speed. arXiv:2212.08059
Ma, Q., Jiang, J., Liu, X., & Ma, J. (2021). Learning a 3D-CNN and transformer prior for hyperspectral image super-resolution. arXiv:2111.13923
Maričić, S., Spaventi, J., Pavičić, L., & Pifat-Mrzljak, G. (1998). Citation context versus the frequency counts of citation histories. *Journal of the American Society for Information Science*, *49*(6), 530–540. https://doi.org/10.1002/(SICI)1097-4571(19980501)49:6<530::AID-ASI5>3.0.CO;2-8
Marshakova-Shaikevich, I. (1996). The standard impact factor as an evaluation tool of science fields and scientific journals. *Scientometrics*, *35*(2), 283–290. https://doi.org/10.1007/BF02018487
Newton, I., & Hooke, R. (1675). Isaac Newton letter to Robert Hooke. https://discover.hsp.org/Record/dc-9792/.
Niveditha, B., Kumbar, M., & Sampath Kumar, B. T. (2022). Rotten web citations cited in scholarly journals: Use of time travel for retrieval. *Aslib Journal of Information Management*, *74*(2), 225–243. https://doi.org/10.1108/AJIM-05-2021-0139
Ott, D. E. (2022). Reference hygiene and death on the internet - decay, rot, half-life, deterioration, and corruption. *JSLS : Journal of the Society of Laparoscopic & Robotic Surgeons*, *26*(1), e2021.00082. https://doi.org/10.4293/JSLS.2021.00082
O'Connor, C., & O'Connor, A. (2018). Reference rot in medical publications. *Irish Medical Journal*, *111*(9), 827.
Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., Moher, D., Yepes-Nuñez, J. J., Urrútia, G., Romero-García, M., & Alonso-Fernández, S. (2021). Declaración PRISMA 2020: Una guía actualizada para la publicación de revisiones sistemáticas. *Revista Española de Cardiología*, *74*(9), 790–799. https://doi.org/10.1016/j.recesp.2021.06.016
Peltonen-Sainio, P., Jauhiainen, L., & Lehtonen, H., et al. (2016). Land use, yield and quality changes of minor field crops: Is there superseded potential to be reinvented in Northern Europe? *PloS one*, *11*(11). QID: Q36199683. https://doi.org/10.1371/journal.pone.0166403
Philibert, A., Loyce, C., & Makowski, D., et al. (2012). Quantifying uncertainties in N2O emission due to N fertilizer application in cultivated areas. *PloS one*, *7*(11). QID: Q28710450. https://doi.org/10.1371/journal.pone.0050950
Radauskas, G. (2024). Laliga accused of blocking legitimate websites amid anti-piracy campaign. Accessed: 2025-05-07. Archived at: https://archive.is/b8xUX https://cybernews.com/news/spain-laliga-streaming-piracy-campaign.
Rosenheim, J. A., & Gratton, C. (2017). Ecoinformatics (big data) for agricultural entomology: Pitfalls, progress, and promise. In M. R. Berenbaum (Ed.), *Annual review of entomology, vol 62* (pp. 399–417). Annual Reviews (*vol. 62*). https://doi.org/10.1146/annurev-ento-031616-035444
Rousseau, R. (1987). The Gozinto theorem: Using citations to determine influences on a scientific publication. *Scientometrics*, *11*(3–4), 217–229. https://doi.org/10.1007/BF02016593
Sahneh, F. D., & Scoglio, C. (2011). Epidemic spread in human networks. Version Number: 1 https://doi.org/10.48550/ARXIV.1107.2464
Saier, T., & Färber, M. (2020). unarXive: A large scholarly data set with publications' full-text, annotated in-text citations, and links to metadata. *Scientometrics*, *125*(3), 3085–3108. https://doi.org/10.1007/s11192-020-03382-z
Saier, T., Krause, J., & Färber, M. (2023a). unarXive 2022: All arXiv publications pre-processed for nlp, including structured full-text and citation network. Publisher: arXiv Version Number: 1 https://doi.org/10.48550/ARXIV.2303.14957
Saier, T., Krause, J., & Färber, M. (2023b). unarXive: All arXiv publications pre-processed for NLP, including structured full-text and citation network (open subset). https://doi.org/10.5281/ZENODO.7752615
Sampath Kumar, B. T., & Vinay Kumar, D. (2013). HTTP 404-Page (not) found: Recovery of decayed URL citations. *Journal of Informetrics*, *7*(1), 145–157. https://doi.org/10.1016/j.joi.2012.09.007
Schmidt, M. (2024). Why do some retracted articles continue to get cited? *Scientometrics*, *129*(12), 7535–7563. https://doi.org/10.1007/s11192-024-05147-4
Sife, A. S., & Lwoga, E. T. (2017). Retrieving vanished web references in health science journals in East Africa. *Information and Learning Science*, *118*(7/8), 385–392. https://doi.org/10.1108/ILS-04-2017-0030
Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science*, *50*(9), 799–813. https://doi.org/10.1002/(SICI)1097-4571(1999)50:9<799::AID-ASI9>3.0.CO;2-G
Stall, S., Yarmey, L., Cutcher-Gershenfeld, J., Hanson, B., Lehnert, K., Nosek, B., Parsons, M., Robinson, E., & Wyborn, L. (2019). Make scientific data FAIR. *Nature*, *570*(7759), 27–29.
Teixeira Da Silva, J. A., & Nazarovets, M. (2023). Archiving website-based references in academic papers: Problems caused by reference rot, potential solutions and limitations. *Learned Publishing*, *36*(3), 477–487. https://doi.org/10.1002/leap.1560
Thelwall, M. (2019). Should citations be counted separately from each originating section? *Journal of Informetrics*, *13*(2), 658–678. https://doi.org/10.1016/j.joi.2019.03.009
Thorp, A. W., & Brown, L. (2007). Accessibility of internet references in annals of emergency medicine: Is it time to require archiving? *Annals of Emergency Medicine*, *50*(2), 188–192.e33. https://doi.org/10.1016/j.annemergmed.2006.11.019
Tuli, S., Casale, G., & Jennings, N. R. (2022a). Carol: Confidence-aware resilience model for edge federations. arXiv:2203.07140
Tuli, S., Casale, G., & Jennings, N. R. (2022b). Dragon: Decentralized fault tolerance in edge federations. arXiv:2208.07658
KumarV., & KumarS. (2019). Recouping the missing web citations in library Hi-Tech journal. *The Journal of Indian Library Association*, *55*(4). https://www.ilaindia.net/jila/index.php/jila/article/view/334.
Wagner, C., Gebremichael, M. D., Taylor, M. K., & Soltys, M. J., et al. (2009). Disappearing act: Decay of uniform resource locators in health care management journals. *Journal of the Medical Library Association : JMLA*, *97*(2), 122–130. https://doi.org/10.3163/1536-5050.97.2.009
Wagner, C. R. H., Niles, M. T., & Roy, E. D., et al. (2019). US County-level agricultural crop production typology. *BMC Research Notes*, *12*(1). QID: Q93018849. https://doi.org/10.1186/s13104-019-4594-4
Wren, J. D. (2004). 404 not found: The stability and persistence of URLs published in MEDLINE. *Bioinformatics*, *20*(5), 668–672. https://doi.org/10.1093/bioinformatics/btg465
Wren, J. D., Johnson, K. R., Crockett, D. M., Heilig, L. F., Schilling, L. M., & Dellavalle, R. P. (2006). Uniform resource locator decay in dermatology journals: Author attitudes and preservation practices. *Archives of Dermatology*, *142*(9). https://doi.org/10.1001/archderm.142.9.1147
Wu, Y., & Ma, W. (2022). Rural workplace sustainable development of smart rural governance workplace platform for efficient enterprise performances. *Journal of Environmental and Public Health*, *2022*. https://doi.org/10.1155/2022/1588638
Yarnelle, L. (2020). Is free public data worth the cost? https://web.archive.org/web/20240522011735/https://riskspan.com/is-free-public-data-worth-the-cost/.
Yayla, K. (2022). Decay of web references in academic publications: A case of Turkish librarianship journal. *Turk Kutuphaneciligi - Turkish Librarianship*, . https://doi.org/10.24146/tk.1048628
Yu, Z., Lu, C., Tian, H., & Canadell, J. G. (2019). Largely underestimated carbon emission from land use and land cover change in the conterminous United States. *Global Change Biology*, *25*(11), 3741–3752. QID: Q91923727. https://doi.org/10.1111/gcb.14768
Zittrain, J., Albert, K., & Lessig, L. (2014). Perma: Scoping and addressing the problem of link and reference rot in legal citations. *Legal Information Management*, *14*(2), 88–99. https://doi.org/10.1017/S1472669614000255