PNAS Nexus, 2024, 3, pgae155
https://doi.org/10.1093/pnasnexus/pgae155
Advance access publication 7 May 2024
Research Report

Hidden citations obscure true impact in science

Xiangyi Meng [a,b], Onur Varol [a,c] and Albert-László Barabási [a,d,e,*]

[a]Network Science Institute and Department of Physics, Northeastern University, Boston, MA 02115, USA
[b]Department of Physics and Astronomy, Northwestern University, Evanston, IL 60208, USA
[c]Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul 34956, Turkey
[d]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA
[e]Department of Network and Data Science, Central European University, Budapest 1051, Hungary
*To whom correspondence should be addressed: Email: a.barabasi@northeastern.edu
Edited By: Erik Kimbrough

Abstract

References, the mechanism scientists rely on to signal previous knowledge, lately have turned into widely used and misused measures of scientific impact. Yet, when a discovery becomes common knowledge, citations suffer from obliteration by incorporation. This leads to the concept of hidden citation, representing a clear textual credit to a discovery without a reference to the publication embodying it. Here, we rely on unsupervised interpretable machine learning applied to the full text of each paper to systematically identify hidden citations. We find that for influential discoveries hidden citations outnumber citation counts, emerging regardless of publishing venue and discipline. We show that the prevalence of hidden citations is not driven by citation counts, but rather by the degree of the discourse on the topic within the text of the manuscripts, indicating that the more discussed is a discovery, the less visible it is to standard bibliometric analysis. Hidden citations indicate that bibliometric measures offer a limited perspective on quantifying the true impact of a discovery, raising the need to extract knowledge from the full text of the scientific corpus.

Keywords: science of science, hidden citation, latent Dirichlet allocation, foundational paper, catchphrase

Significance Statement

When a discovery or technique becomes common knowledge, its citations suffer from what Robert Merton called "obliteration by incorporation." This phenomenon leads to the concept of hidden citations, representing unambiguous textual references to a discovery without an explicit citation to the corresponding manuscript(s). Previous attempts to detect hidden citations have been limited to manually identifying in-text mentions. Here, we use machine learning to systematically identify hidden citations, finding that they emerge regardless of publishing venue and discipline, their frequency being influenced by the level of discussion within manuscript texts. Hidden citations lead to inevitable credit distortion and capture the "burden" of success in science: the more widely a concept is used, the more hidden it is from standard bibliometric analysis.

Introduction

"We stand on the shoulders of giants," the oft-quoted statement acknowledging the cumulative nature of knowledge, has an explicit carrier in the contemporary scientific discourse: the *citation*. Since the 1960s, references—which serve primarily as a mechanism to signal prior knowledge, enhance credibility, and protect against plagiarism—have taken on a secondary role of *allocating scientific credit*, turning into an often used and misused measure of scientific impact (1–3). Yet, when a discovery or technique becomes common knowledge to such a degree that it does not warrant citation any longer, citations suffer from what Robert Merton in 1968 called "obliteration by incorporation (OBI) (4, 5)." For example, concepts like general relativity or black hole evaporation today are so embedded into scientific literacy, that only rarely do manuscripts focusing on the topics cite Einstein's 1915 work (6) or Unruh's 1976 paper (7). As a consequence, foundational ideas of science are undercited, without being underused. This phenomenon leads to hidden citations, representing unambiguous allusions to a body of knowledge without an explicit citation to the manuscript(s) that introduced it. Hidden citations, also known as implicit, indirect (8) or informal citations (9), can also be induced by restrictions imposed by publishing venues on the number of references, prompting authors to cite reviews and books to signal a wider body of knowledge, rather than crediting the original discoveries. While Merton considered such hidden citations the highest level of acknowledgement—a badge of honor,

Competing Interest: A.-L.B. is co-scientific founder of and is supported by Scipher Medicine, Inc., which applies network medicine strategies to biomarker development and personalized drug selection.
Received: February 6, 2024. **Accepted:** April 2, 2024

rather than a negative effect (4), such credit is no longer accessible to traditional bibliometric measures. In the era when citations are widely used as measures of impact (10–18), hidden citations remain *hidden*—not to human beings—but to the quantitative and statistical tools frequently used to quantify scientific credit. This leads to systematic distortion of the credit landscape, diminishing the quantifiable impact of the very discoveries that define scientific progress.

Previous attempts to detect hidden citations have been limited to manually searching and identifying in-text mentions such as "Southern blot (8)," "density functional theory (9)," "Nash equilibrium (19)," "evolutionarily stable strategy (20)," and "bounded rationality (21, 22)." Yet, the lack of automated methodology for determining in-text allusions [including eponyms (23, 24), relating a person to a discovery] and their corresponding primordial references (25) has limited our ability to understand the prevalence of hidden citations to a narrow corpus of manually inspected papers, raising the need for "a more comprehensive estimate of uncitedness (8)." Here, we fill this gap by using machine learning to automatically detect *catchphrases* (19), representing in-text allusions to specific discoveries, matching them with the appropriate primordial references called *foundational papers*. The method allows us to systematically identify hidden citations across the whole scientific literature, and to trace the factors responsible for credit distortion. As roughly 90% of obliteration by incorporation happens in the main text of a manuscript (21), we apply machine learning to the full text, helping us better capturing the accumulation and distortion of credit in science.

Results

Each scientific discovery builds on a body of knowledge embodied by latent topics that are topically named within a manuscript and accompanied by citations to the foundational papers. For example, papers focusing on anti-de Sitter/conformal field theory (AdS/CFT), exploring the correspondence between general relativity and quantum field theory, cite the 1999 paper that introduced the concept (Fig. 1a). Yet, many papers on AdS/CFT use language that for experts unambiguously defines the paper's topic, without citing the foundational work. To identify such hidden citations, we use the Latent Dirichlet Allocation (LDA) model (27, 28) to detect topics in the text of a publication, inferring latent topical structures from a corpus of full-text citation contexts based on symbolic natural language processing and Bayesian inference. In contrast with neural-network-based Word2Vec (29) or BERT (30), the LDA model is an unsupervised machine learning approach that is interpretable, allowing us to associate the outcomes of LDA with confidence levels through transparent probabilistic logic (see Materials and methods).

We identified 343 topics in physics that accumulate hidden citations, each with at least one catchphrase and at least one foundational paper (see Materials and methods). Shown as examples are four topics uncovered by the algorithm (Fig. 1b), as well as the followers of each topic (Fig. 1c–f), defined as papers that either cited the foundational papers of the given topic or mentioned the corresponding topic-specific catchphrases, or both (see Materials and methods). For example, the orange regions denote the temporal evolution of the number of papers that simultaneously cite the foundational papers and carry the respective catchphrases. The top green region captures hidden citations, counting papers that make an unambiguous textual reference to the topic but fail to cite any of the foundational papers. For example, less than half of the papers that use the catchphrases of "AdS/CFT" cited any of the two foundational papers: 1999 paper by Maldacena and 1998 paper by Gubser *et al.* (Fig. 1c). Taken together, we find that for the four topics featured in Fig. 1(b), hidden citations correspond to 65.8%, 61.7%, 34.6%, and 52.3% of all detectable credit since the publication of the respective topic's first foundational paper, overcoming the bibliometrically quantifiable and tabulated citations.

The high proportion of hidden citations prompts us to calculate the temporal changes in the conditional probability that a paper that mentions the topic-specific catchphrases cites the foundational papers, $p(\text{cite}\,|\,\text{mention})$ (Section S9). We find that the probability that the foundational papers are cited drops by approximately 20% after 20 years (Fig. 2a), indicating that the reliance on hidden citations, hence OBI, strengthens over time.

Do hidden citations correspond to pure untracked credit [a.k.a. implicit citations (8)], or is credit diverted to other works [a.k.a. indirect citations (8)]? To distinguish these two mechanisms, we identified the most frequently cocited publications accompanying a hidden citation. We find that for "AdS/CFT" the most cited alternative is a review coauthored by the authors of the two foundational papers, and for "DMRG" the most cited alternatives are two books (Fig. 3a). Credit is also diverted to applications of the topic, such as the application of AdS/CFT to topological quantum field theory, or to extensions on the topic, like in the expanded "BOSS" datasets (Fig. 3a). Overall, we find that the works that collect the credit from hidden citations tend to cite the foundational papers, or cite papers that in turn cite the foundational papers (Fig. 3b). Indeed, around 60% of hidden citations have a citation path length of 2 to the foundational papers (Fig. 3c–f). indicating that hidden citations do cite and give credit to papers whose topics closely relate to the foundational papers. To determine whether the previously observed increase in reliance on hidden citations over time (Fig. 2a) is dominated by implicit citations or indirect citations, we recalculated the temporal changes (Fig. 2a), this time also including indirect citations, i.e. the hidden citations that have a citation path length of at most 2. We find now that $p[(\text{cite} + \text{indirectly cite})\,|\,\text{mention}]$ increases with time (Section S10), indicating that the increasing reliance on hidden citations is accompanied by an increasing tendency to divert credit to other works.

As Fig. 2(b) shows, topics with more citations (c) tend to accumulate more hidden citations (h), a trend approximated by a sublinear dependence $h \sim c^{0.763}$, indicating that on average a topic with 5,000 citations accumulates approximately 1,000 hidden citations. While this scaling suggests that citations are the main driving force of hidden citations, our measurements indicate otherwise. Indeed, we find a negligible correlation ($\rho \approx 0.016$) between $p(\text{cite}\,|\,\text{mention})$ (the probability of being cited if mentioned) and the number of citations for the respective topic (Fig. 2c) (Section S8). We find, however, a strong negative correlation ($\rho \approx -0.611$) between $p(\text{cite}\,|\,\text{mention})$ and the number of mentions per topic (Fig. 2d). In other words, the more discussed is a discovery in the textual context of a paper, the less likely scientists feel the need to explicitly cite it, a "burden" of success that is independent of the publication venue (Fig. 2e–h).

To explore the impact of hidden citations on bibliometric measures, it is tempting to calculate the hidden citations of *individual* foundational papers (Section S11). We must approach this with caution: since our methodology operates at the topic level, transferring citation counts from topics down to individual papers is inherently imprecise and cannot be guaranteed to be accurate. Consequently, the paper-specific observations we offer here should be viewed as provisional insights rather than definitive conclusions.

Our first observation is that the ratio of hidden to explicit citations is, on average, 0.98:1, indicating that papers tend to acquire

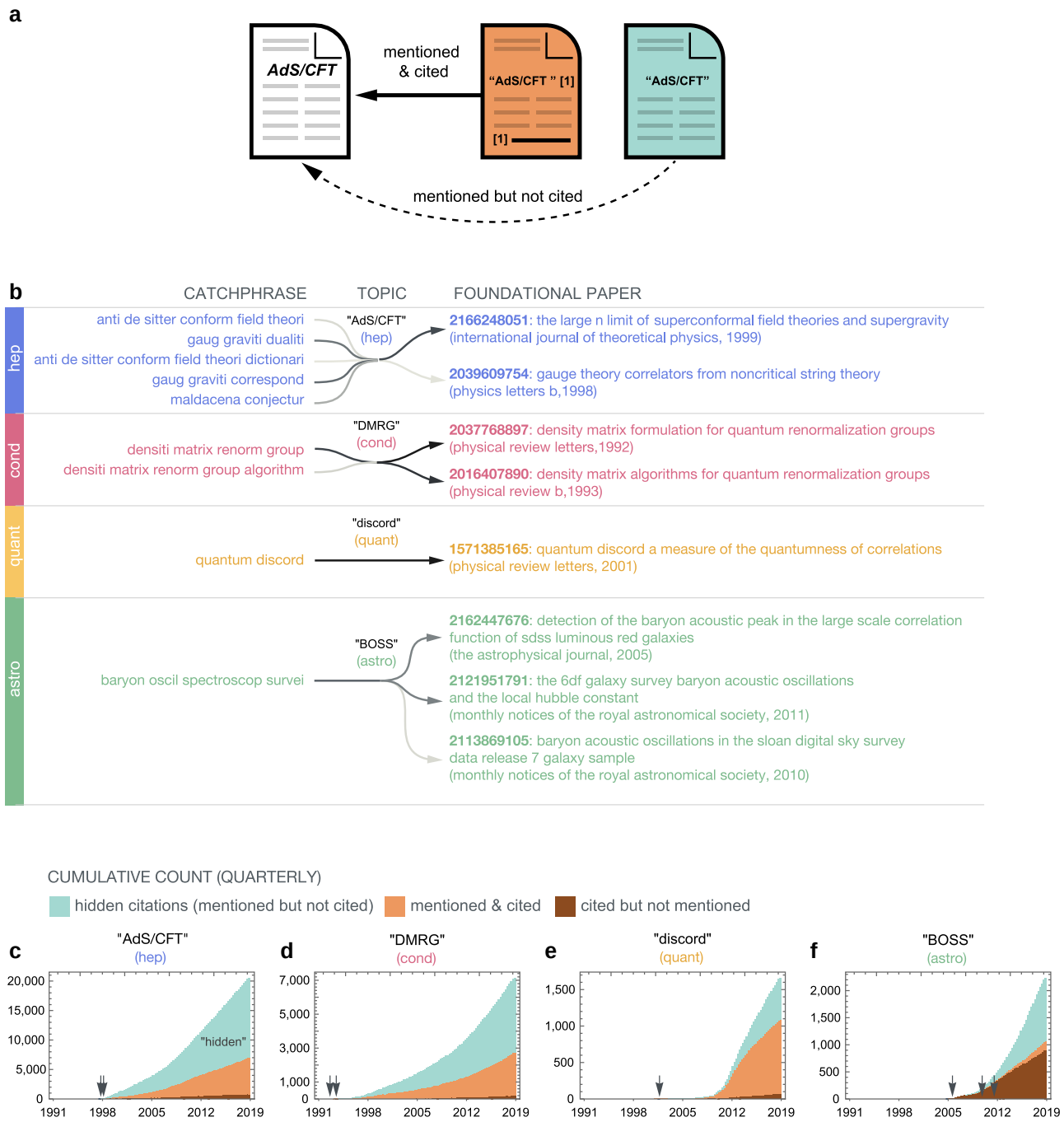

Fig. 1. Hidden citations. a) A foundational paper is a manuscript that introduces a new concept that subsequently defines a topic of inquiry by the scientific community, such as the topic "anti-de Sitter/conformal field theory," also known as "AdS/CFT (26)." Papers focusing on the topic mention the catchphrase "AdS/CFT" or "anti-de Sitter/conformal field theory," followed by a citation to one of the foundational papers. Often, however, the catchphrases are present without explicit citations, resulting in hidden citations. b) Exemplary topics selected from high energy physics (hep), condensed matter physics (cond), quantum physics (quant), and astrophysics (astro), together with their corresponding catchphrase(s) (lemmatized as word stems) and foundational paper(s) (Microsoft Academic Graph id). Darker arrows denote the algorithm's higher statistical confidence for the respective foundational paper. c–f) Time evolution of citations and hidden citations for the topics listed in (b). The arrows denote the publication date(s) of the foundational paper(s) for each topic.

hidden citations at the same rate as they acquire explicit citations. Yet, we do observe considerable variability in this ratio. Hence, for some foundational papers hidden citations can dominate over explicit citations. Examples include the paper introducing the cosmological inflation theory in 1981 that acquired 8.8 times more hidden citations than explicit citations, or the 1974 work that merged the electromagnetic, weak, and strong forces into a single force, which accumulated 6.6 times more hidden citations than explicit citations. This prompted us to calculate the changes in citation-based ranks between foundational papers (Section S11). As Fig. 4(a) indicates, most papers in the top 100 list suffer rank loss (green lines), thanks to a few publications that accumulate an exceptional number of hidden citations, and gain significantly in rank (red lines). For example, the most cited paper of arXiv, the 1999 paper which started the formal theory of AdS/CFT, loses its top ranking once we take hidden citations into

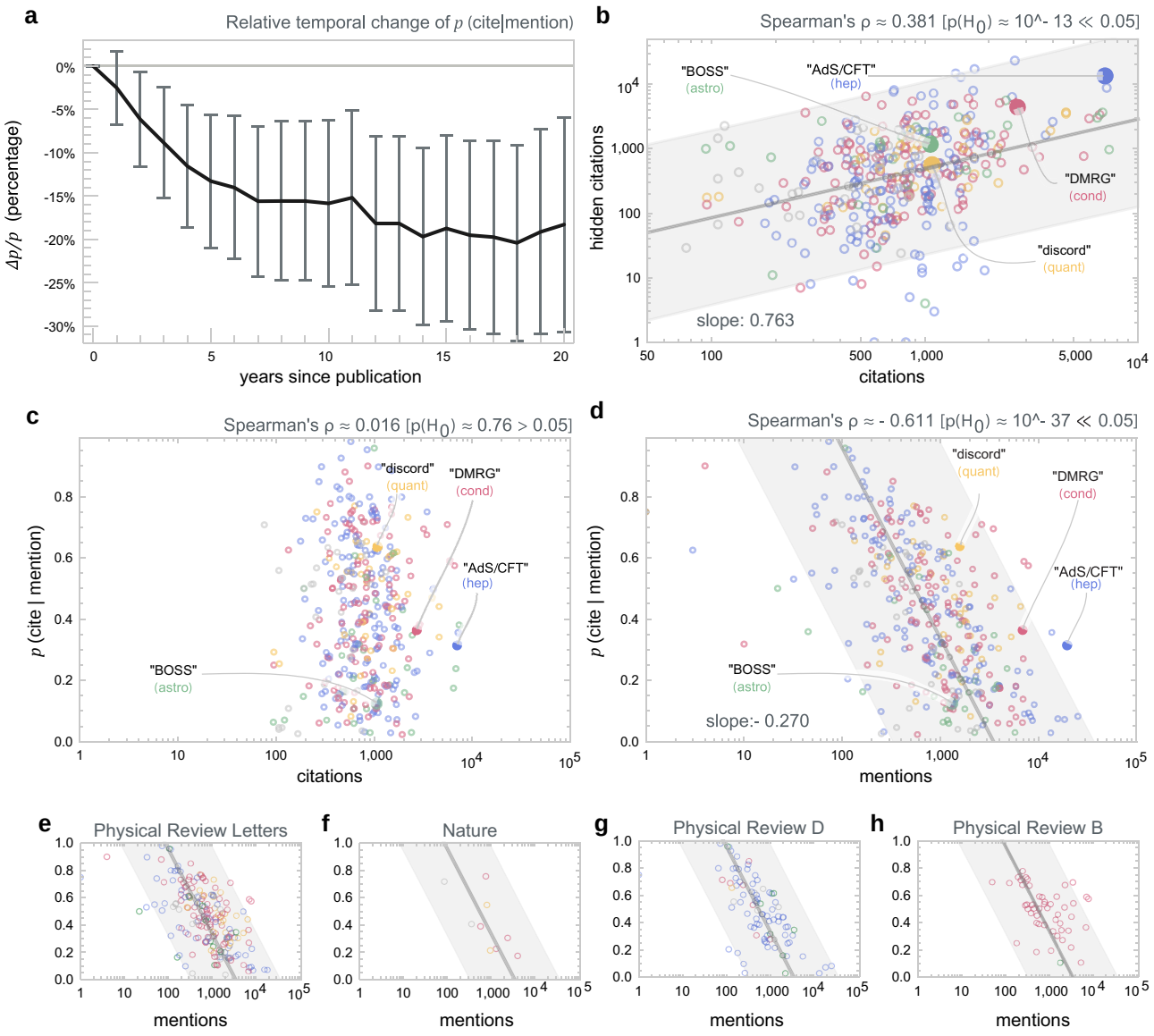

Fig. 2. Factors that drive hidden citations. a) The temporal change of p(cite | mention), the probability that a paper mentioning the topic-specific catchphrases will also cite the foundational paper, as a function of time (years since publication). On average, p(cite | mention) per topic drops by approximately 20% after 20 years of publication of the first foundational paper. Error bars represent 95% confidence intervals. b) Topics with more citations (c) tend to have more hidden citations (h) (with Spearman's rank correlation $\rho \approx 0.381$ and null hypothesis H_0 rejected). Most topics fall into the 95% single-observation confidence bands with a log–log slope 0.763±0.208, indicating that $h \sim c^{0.763}$. c) p(cite | mention) as a function of citations per topic ($\rho \approx 0.016$, H_0 not rejected), indicating that the probability of a textual reference becoming a hidden citation is not driven by the number of citations to the topic. d) p(cite | mention) as a function of mentions per topic ($\rho \approx -0.611$, H_0 rejected). The strong negative correlation indicates that hidden citations are driven by the number of textual mentions of the topic. Most topics fall into the 95% confidence bands with a log-linear slope −0.27±0.04. The pattern holds for four distinct publication venues e–h).

account, to the 1981 paper previously ranked #8, which started the phenomenological study of the cosmological inflation theory (Fig. 4a). Hidden citations could potentially have an impact on authors as well. To see if this is the case, we adopted the Microsoft Academic Graph's "author saliency" metric, that relies on the heterogeneous network structure of the connectivity of articles, authors, and journals, designed to be less susceptible to raw citation counts and temporal bias (31). We find that, when we compare two foundational papers with similar numbers of explicit citations, authors with more hidden citations have higher average saliency, a positive correlation notable for papers with less than 3,000 citations (Fig. 4b). This suggests that OBI tends to correlate with a *positive* impact on authors whose papers' full citation impact has not yet developed. Interestingly, the effect disappears for well-recognized papers, for which missing citations does not appear to affect their authors' reputation.

Papers that became foundational papers and acquired hidden citations tend to be highly cited, accumulating on average 434±34 explicit citations, in contrast with 1.4 explicit citations for all physics papers in the corpus. Yet, not only highly cited papers acquire hidden citations. We find that even among papers with citations $\lesssim$500, a nonnegligible fraction (>10%) of papers may acquire hidden citations (Fig. 4c). Since our approach to identifying hidden citations (see Materials and methods) is conservative, designed to reduce false positive errors, the actual fraction of papers that acquire hidden citations is likely higher.

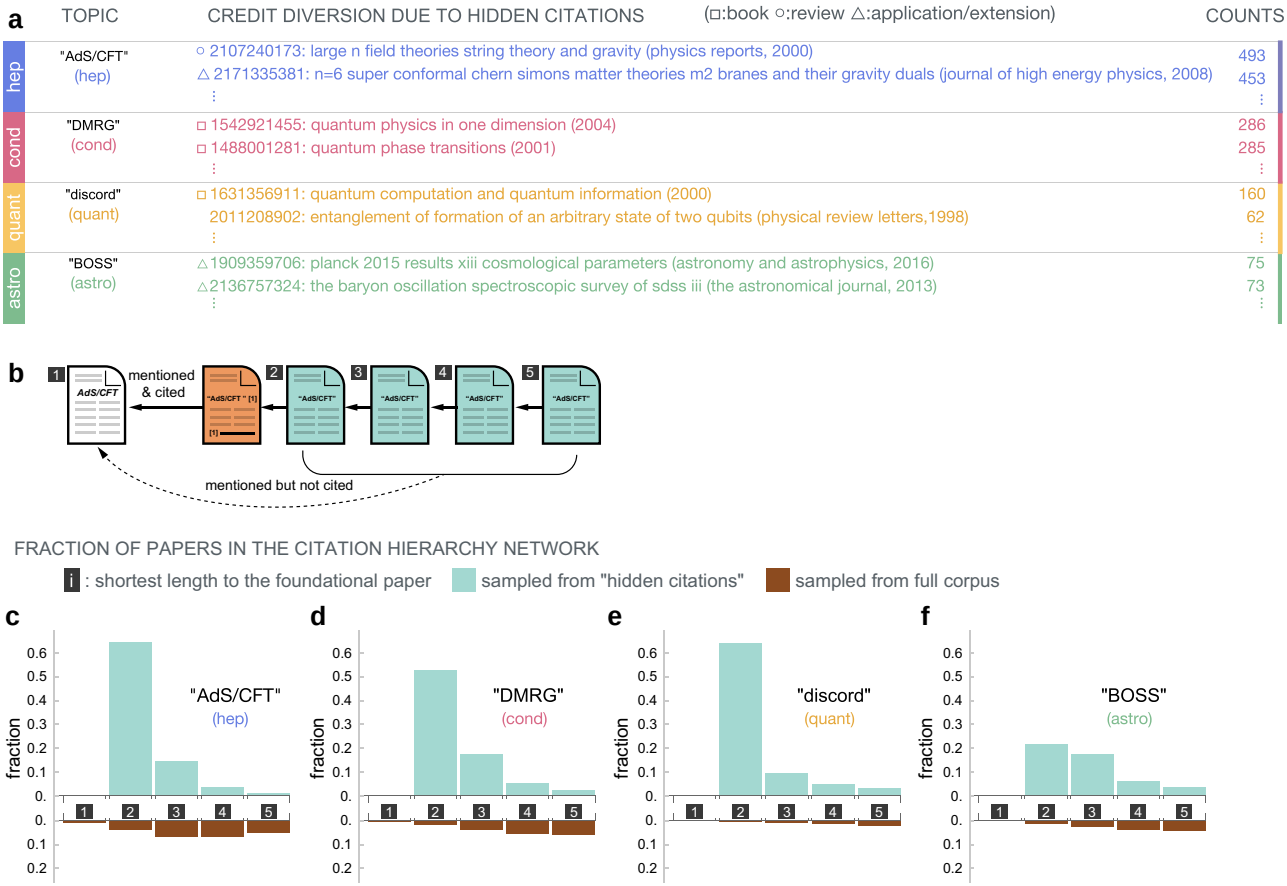

Fig. 3. Credit redirected. a) The most cited alternatives for four topics that acquire hidden citations, primarily indicating that credit is often diverted to books, reviews or applications/extensions of the foundational papers. b) Most alternatives to hidden citations are related to the foundational paper, detectable by tracking the citation path between the alternative and the foundational paper. c–f) Fraction of hidden citations ranked by their citation hierarchy to the foundational papers. For each topic (except "BOSS"), around 60% of hidden citations (green, top) cited other arXiv papers that explicitly cited the foundational papers. For a randomly sampled reference from the full arXiv, this fraction is negligible (brown, bottom).

We also find that hidden citations are not limited to older papers, but they accompany recent publications as well (Fig. 4d), such as the discovery of gravitational wave (2016) or exclusion of dark matter particles in the Large Underground Xenon experiment (2017).

Finally, we investigated the sociodemographic characteristics (gender, country of origin, and the prestige of institution) of the authors of foundational papers (Section S12). We find that hidden citations capture a universal phenomenon that can emerge in any institution, regardless of their level of prestige, and we observe no statistically significant bias based on gender or country of origin in hidden citations apart from the overall biases in explicit citations that have long been observed (32–35). Hence, although the methodology presented in the paper can account for additional hidden citations, it simply reflects the existing biases.

Depending on the specific assumptions made in our methodology to redistribute credit from topics to individual papers, the observations (Fig. 4a–d) will likely look different. This serves as a reminder of the significant role hidden citations—and how they are calculated—play in the allocation of scientific credit.

These observations highlight the potential risks associated with ignoring or misidentifying hidden citations in credit allocation. Such risks raise an important question: what determines the emergence of hidden citations? Our analysis suggests that there is a prerequisite for a paper to acquire hidden citations: it must develop exclusive catchphrases that are synonymous to the paper itself, becoming a "conceptual symbol" (36, 37) within the field. For example, whenever "quantum discord" is mentioned, an expert in the field will immediately link it to the 2001 foundational paper, and vice versa, seeing a citation to that particular reference, an expert thinks of "quantum discord." To quantify this dual correspondence, we first measured the degree of nonexclusivity of linking a given n-gram w to a paper d by calculating the specific conditional entropy $\mathcal{S}(d\,|\,w)$ (Section S13), finding that $\mathcal{S}(d\,|\,w)$ is considerably lower for catchphrases than for non-catchphrases. For example, $\mathcal{S}(d\,|\,''\text{quantum discord}'') \approx 4.07$ in contrast with $\mathcal{S}(d\,|\,''\text{quantum mechanics}'') \approx 6.73$, indicating that "quantum discord" is a catchphrase pointing to a well-defined foundational paper, while "quantum mechanics" is too general to be exclusively assigned to one or a few foundational papers. Inversely, we measured the specific conditional entropy $\mathcal{S}(w\,|\,d)$ of seeing paper d and linking it to an n-gram w, finding again that $\mathcal{S}(w\,|\,1571385165) \approx 5.97$ for the 2001 foundational paper with catchphrase "quantum discord" is lower than the highly cited 1999 paper also focusing on quantum information processing, $\mathcal{S}(w\,|\,2097039598) \approx 7.72$, but not categorized by our algorithm as a foundational paper. These results confirm that to develop hidden citations, a (catchphrase)—(foundational paper) pair must acquire *mutual* exclusivity: a paper does not accumulate hidden citations if its conceptual significance does not lead to an unambiguous catchphrase, or if that catchphrase is not exclusive enough for the community to unambiguously link it back to the original paper.

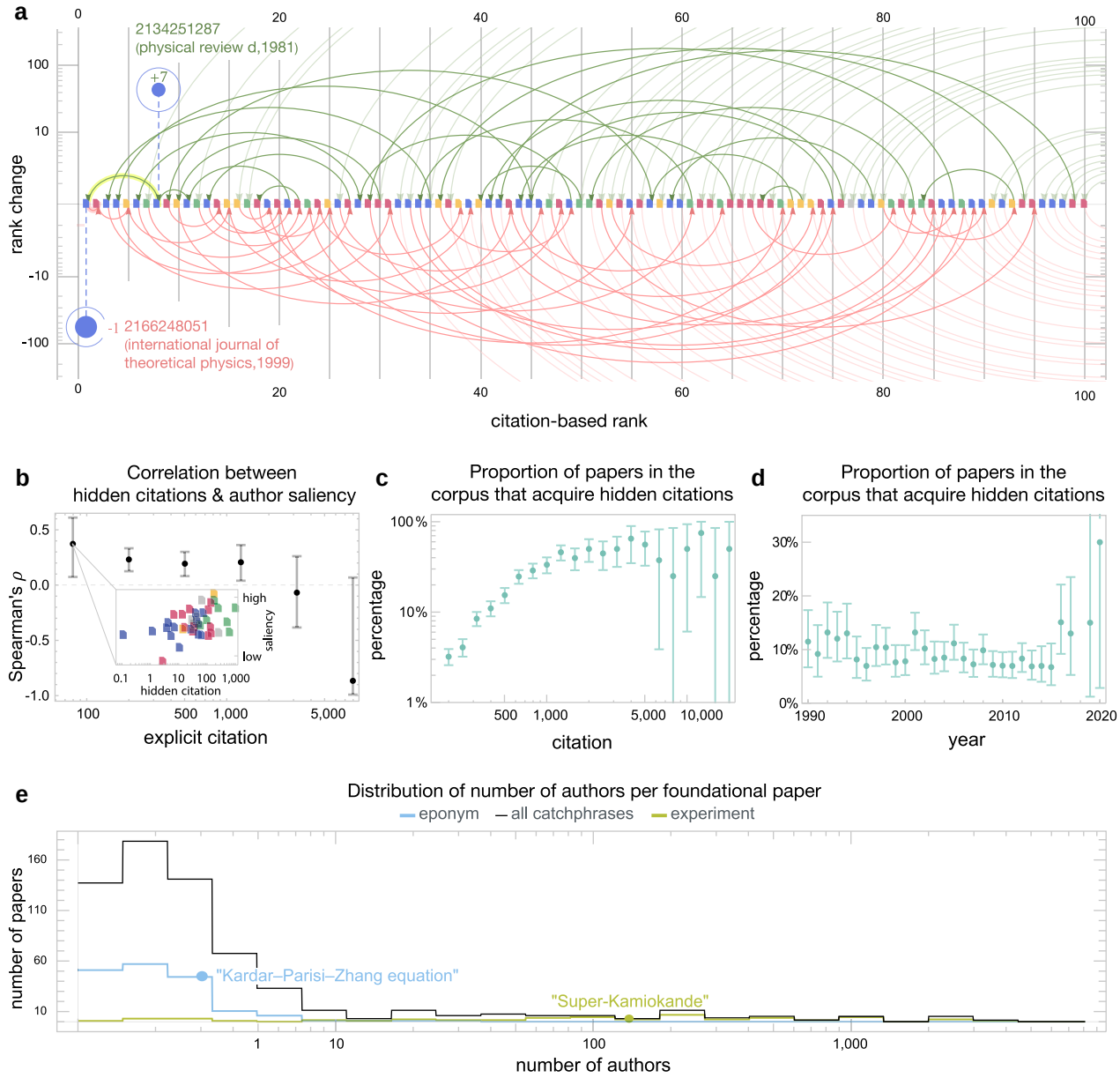

Fig. 4. Foundational papers. a) Changes in the citation-based ranks of the top-ranked foundational papers after taking hidden citations into account, shown by arrows from the old explicit-citation-based rank to the new explicit-plus-hidden-citation-based rank (green: rank rise; red: rank drop). After accounting for hidden citations, the "cosmological inflation theory" paper (2134251287), ranked #8 based on explicit citation counts, takes the top spot. b) For foundational papers with similar numbers of explicit citations, the paper with more hidden citations tends to result in higher average author saliency (inset). The proportion of papers in the corpus that can acquire hidden citations increases with c) the explicit citations but not with d) the publication year of the papers. Error bars represent 95% confidence intervals. e) Distribution of foundational papers by the number of authors per foundational paper, shown for all catchphrases (black) and for eponym-related (blue) and experiment-related catchphrases (green).

Asking where such exclusive catchphrases originate, we find that for 78.8% of the 880 foundational papers the corresponding catchphrases do not appear in the titles or the abstracts of them (Section S14). This indicates that catchphrases are typically not proposed by the authors of the foundational papers, but are assigned later by the community (37). We also find that 26.0% of all foundational papers have catchphrases that correspond to eponyms (e.g. "Kardar–Parisi–Zhang equation," that governs surface growth) and another 7.1% acquire the names of experimental projects (e.g. "Super-Kamiokande," the discovery of neutrino oscillation) (Section S14). Eponym-related catchphrases emerge mainly for papers with short author lists—indeed, foundational papers with eponyms as catchphrases have 2.50±0.36 authors on average, in contrast with 72.4±26.5 authors for noneponym-related catchphrases, and 405±164 for papers with experiment-related catchphrases (Fig. 4e).

Identifying hidden citations in all areas of science requires a large and unbiased corpus of full-text citation contexts. While such a corpus is so far unavailable for all science, we have access to 818,311 computer science and 140,865 biology full-text manuscripts, allowing us to identify catchphrases and foundational papers in these fields as well (Fig. 5a). The patterns governing hidden citations are largely indistinguishable from those documented for physics: we observe a significant number of hidden citations for established research topics like "Kalman" (refining estimates from new measurements) and

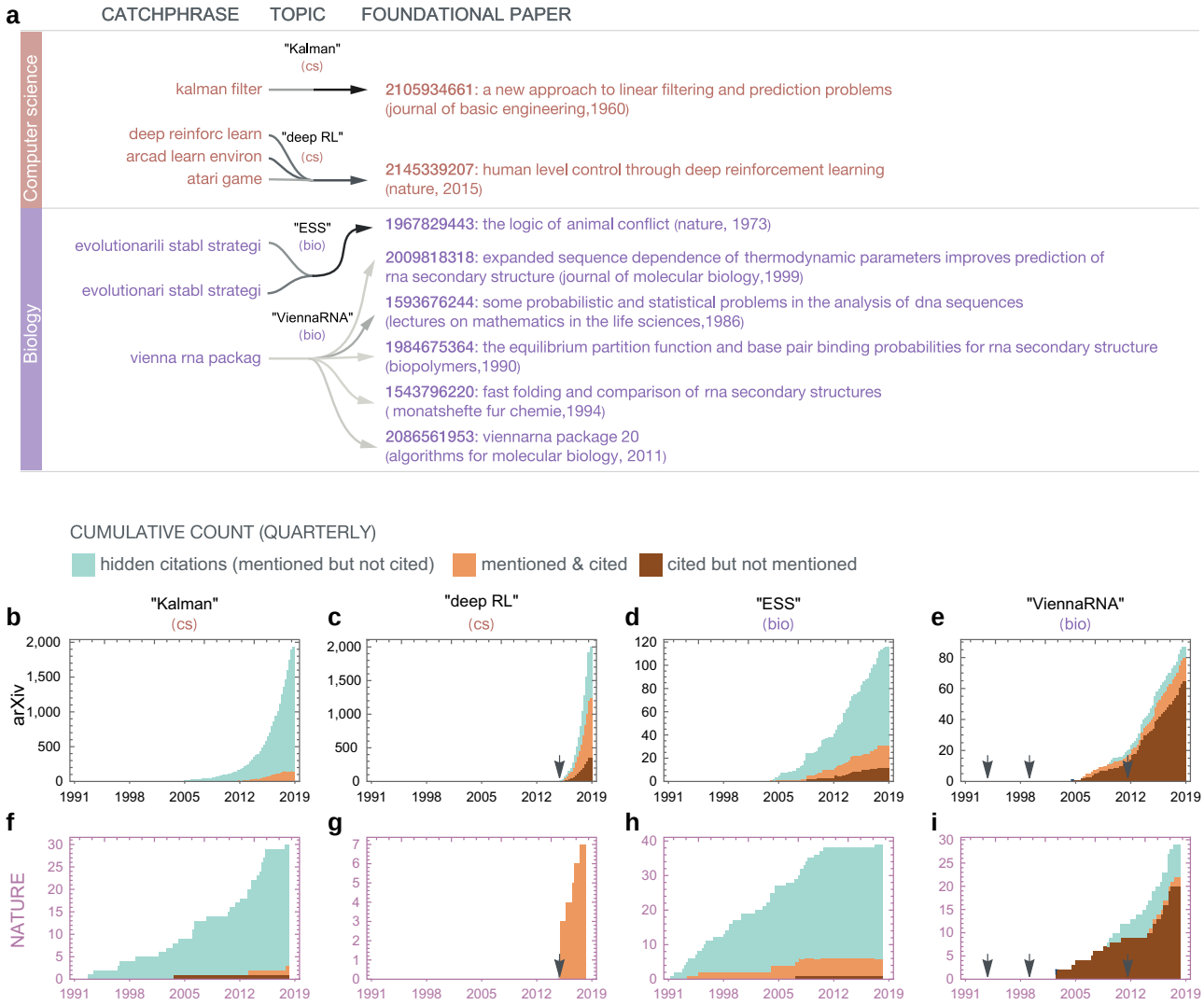

Fig. 5. Hidden citations across disciplines and venues. a) Four topics selected from computer science (cs) and biology (bio) (cf. Fig. 1b). b–i) Time evolution of citations and hidden citations (cf. Fig. 1c–f) for the four topics shown in a), identified from arXiv b–e) and *Nature* f–i).

"ESS" (evolutionary strategies in natural selection) and even for newer topics like "deep RL" (deep neural networks and artificial intelligence) and "ViennaRNA" (analysis of RNA structures) (Fig. 5b–e). We also analyzed a corpus of 88,637 full-text *Nature* articles (38), which cover multiple disciplines, finding evidence of hidden citations in highly selective peer-reviewed venues as well (Fig. 5f–i). These results indicate that hidden citations are a universal phenomenon, emerging in all areas of science and publishing venues, disciplines, and research topics.

Discussion

Acknowledging discoveries on which new research builds on is an integral part of the scientific discourse. Yet, with the exponential growth of science and limits on the number of allowed references, a paper's ability to credit all sources of inspiration is limited. Such limitations lead to inevitable credit distortion, manifest in situations where the textual context indicates that credit is due, but it is not accompanied by explicit citations to the pertinent work. Hidden citations capture the "burden" of success in science: the more widely a concept is used by the scientific community, the more likely that it will accrue hidden citations.

Systematically tabulating hidden citations, together with explicit citations can help us more accurately identify emerging topics and evaluate their true impact (8, 9). That being said, both explicit and hidden citations represent unequal "atoms of peer recognition (39)," offering different degrees of credit per citation. Indeed, when citations point to highly cited papers (driven by the authors' fame, race, gender, etc.) without contributing to the paper's topic, they offer less credit (40). Negative citations (41) should also offer less credit than positive citations; yet, we find that the prevalence of negative or positive texts is low in both the explicit and hidden citations of foundational papers (Section S6). This can be attributed to the fact that in order for a paper to become a foundational paper and acquire hidden citations, the discovery or technique it presents should have already been accepted as common knowledge, leaving limited room for debates. There is also a difference in crediting conceptual versus methodological advances: while papers that cite the foundational papers but fail to give appropriate textual references are extremely rare (Figs. 1c–e and 5b–d, f–h), for papers introducing databases

(Fig. 1f) or tools (Fig. 5e,i) textual references are less frequent. One explanation is that these foundational papers are not just cited for their dataset or methodological efforts; they are also frequently cited for supporting the corresponding general concept, namely, "baryon oscillation" or "RNA structure." In the latter case, authors often fail to mention the words "survey" or "package" when citing these papers. This textual bias suggests that database or methodological advances are often less acknowledged, in line with earlier findings (42). This is because when playing the supportive role of a general concept, the papers lost their merits as foundational papers. The community working on the general concept often benefits from the database or methodological efforts without textually referencing and explicitly acknowledging the effort that went into creating it (Section S7).

It is, therefore, important to go beyond simple counts of explicit and hidden citations, and develop new metrics that can also differentiate the degree of credit carried by each citation, a process to which a complete corpus of both explicit and hidden citations is a prerequisite.

While our unsupervised methodology allows us to tabulate hidden citations at scale, the current methodology is designed to be conservative and to minimize false positive errors (see Materials and methods). Thus, currently it may overlook hidden citations, limiting the completeness of topics. Note that the missing hidden citations can be recovered by lowering the identification thresholds, at the expense of increasing false positive errors. Another limitation is that some topics may not be present in the arXiv corpus, either because they have not been studied or discussed in a sufficient number of arXiv papers or because they are too narrow or outdated. They could be recovered if we apply our algorithm to a more extensive full-text database that spans multiple disciplines and time periods. However, there is a major barrier to achieving this: the lack of systematic access to full-text papers. Indeed, while citation counts and other metadata are now freely and easily available for research purposes, access to the full text of all research papers is restricted by commercial interests, limiting the deployment of tools capable of accurately tabulating hidden citations and their role in the scientific discourse.

Materials and methods

Traditionally, the LDA model is used to uncover latent topics within a collection of documents. Each document is assumed to be a mixture of multiple topics, and each topic is characterized by a distribution over phrases. Here, instead of exploring latent topical structures, we focus on the explicit textual observables, aiming to reveal the correspondence between phrases and documents.

The input of the LDA model is a list of 2-tuples $\{w, d\}$ between an n-gram w (a phrase of n words, where the value of n can freely vary to accommodate long phrases) and an accompanied text-based *document* d (denoted by a unique code, e.g. the Microsoft Academic Graph [MAG] id (43)). The document does not contain the full text of the MAG paper d. Instead, it comprises the citation contexts of d, which represent the textual discussions by the community when citing d. Each 2-tuple accounts for an exact occurrence of an n-gram w in document d. For example,

$$\text{input} = \begin{pmatrix} \{''\text{string theory}'', 2166248051\} \\ \{''\text{gauge-gravity duality}'', 2166248051\} \\ \{''\text{quantum discord}'', 1571385165\} \\ \cdots \end{pmatrix}.$$

The output of the LDA model is a list (of the same length of the input) of 3-tuples $\{w, z, d\}$, where each input 2-tuple $\{w, d\}$ acquires a new latent variable z that corresponds to a specific topic, such as

$$\text{output} = \begin{pmatrix} \{''\text{string theory}'', \text{topic 1}, 2166248051\} \\ \{''\text{gauge-gravity duality}'', \text{topic 1}, 2166248051\} \\ \{''\text{quantum discord}'', \text{topic 2}, 1571385165\} \\ \cdots \end{pmatrix}.$$

The output indicates that "string theory" and "gauge-gravity duality" belong to the same topic 1, while "quantum discord" belongs to a different topic 2. The joint probability $P(w, z, d)$ of the concurrent occurrence of the 3-tuple $\{w, z, d\}$, which can be estimated from the output, enables us to define and calculate two key terms:

1. An n-gram w with $P(z \mid w) > P_{\text{th}}^{\text{catch}}$ is a *catchphrase* of topic z, implying that whenever the n-gram w is seen in a document, we are confident that topic z is referred to. For example, if there are 1,919 occurrences of $\{w =''$ quantum discord$'', z, d\}$ in the output, among which 1,916 3-tuples also have $z = 3$, then $P(z \mid w) \approx 0.998 \pm 0.002$, representing the conditional probability of referring to topic $z = 3$ given occurrence of $w =$"quantum discord." If $P(z \mid w)$ is larger than $P_{\text{th}}^{\text{catch}}$, then we have statistical confidence that "quantum discord" is a catchphrase of the topic $z = 3$.
2. A document d with $P(d \mid z) > P_{\text{th}}^{\text{found}}$ is a *foundational paper* of topic z, implying that whenever topic z is referred to, we expect a citation to the MAG paper d, indicating that the foundational paper is sufficiently disruptive (16) to serve as a representative of the topic. For example, if there are 9,742 occurrences of $\{w, z = 3, d\}$ in total in the output, among which 2,091 3-tuples also include the document $d = 1571385165$, then $P(d \mid z) \approx 0.215 \pm 0.008$ which, if larger than $P_{\text{th}}^{\text{found}}$, makes $d = 1571385165$ a foundational paper of topic $z = 3$.

We rely on a strict criterion to choose catchphrases ($P_{\text{th}}^{\text{catch}} = 0.95$) but a loose criterion at including foundational papers ($P_{\text{th}}^{\text{found}} = 0.05$), reducing the false positive rate of incorrectly assigning a too general n-gram as a catchphrase, or concentrating too much hidden citations for only one or two papers, hence remaining conservative at identifying hidden citations per foundational paper. This unavoidably results in the exclusion of some topics for which the catchphrases and foundational papers are less exclusively defined. Therefore, our results are not based on a complete collection but a sampled aggregation of topics.

After a latent topic z is inferred by LDA, we identify all papers that follow and explore the topic z (Fig. 1a), including all the papers that explicitly cite the foundational paper(s) of topic z, as well as papers that only mention the topic-specific catchphrase(s) but lack citations to the foundational paper(s) (Section S4). The latter corresponds to hidden citations, as they explicitly build on the catchphrase(s) associated with topic z. For example, a hidden citation is detected when a paper mentions the catchphrase "quantum discord" but lacks a citation to the foundational paper $d = 1571385165$.

We trained the LDA classifier using the *unarXive* dataset (44) that offers full-text coverage for 1,043,126 publications, annotated with citation contexts, obtained after merging the entire arXiv (45) with MAG (43) (Section S1). Established in 1991 as the first preprint archive, arXiv offers a fairly unbiased coverage of physical sciences. We identified from the citation contexts (from arXiv) all n-grams w and each paper d (in MAG) they refer to

(Section S2), initially filtering out books and reviews. Following the arXiv taxonomy, the results are categorized into five categories (Section S5): high energy physics ("hep"), condensed matter physics ("cond"), quantum physics ("quant"), astrophysics ("astro"), and the rest ("other"). For example, the LDA model predicts that each time the catchphrase "anti-de Sitter conform field theory" is mentioned, it should be accompanied by either a reference to the 1999 paper (2166248051) by Maldacena (26), or to (2039609754) by Gubser, Klebanov, and Polyakov, both within the "hep" topic "AdS/CFT" (Fig. 1b). Similarly, for the "density matrix renormalization group" catchphrase, the LDA model expects references to two papers (2037768897 and 2016407890) by White, within the "cond" topic "DMRG" that focuses on many-body ground-state wave functions (Fig. 1b). To validate the accuracy of LDA, we have consulted specialists in "hep" and "quant" to manually check ten randomly selected topics each (Section S5). We find that 9 out of 10 of the specialists' choices of the foundational papers are identified by the algorithm, resulting in a 90% effectiveness of our automated approach.

Acknowledgments

We would like to thank Alice Grishchenko for help with data visualization. X.M. is indebted to Yanchen Liu, Rodrigo Dorantes-Gilardi, Bingsheng Chen, Alexander J. Gates, Louis M. Shekhtman, and Jing Ma for fruitful discussions. This manuscript was posted on arXiv:2310.16181. A.-L.B. is the scientific founder of Scipher Medicine, Inc., which applies network medicine to biomarker development, of Foodome, Inc., which applies data science to health, and of Datapolis, Inc., which focuses on human mobility.

Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

Funding

This research was funded by the National Science Foundation (SES-2219575), the Eric and Wendy Schmidt Fund for Strategic Innovation (G-22-63228), John Templeton Foundation (#62452), and Air Force Office of Scientific Research (FA9550-19-1-0354). X.M. was supported by the NetSeed: Seedling Research Award of Northeastern University. A.-L.B. is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 810115–DYNASNET.

Author Contributions

All authors designed and did the research. O.V. and A.-L.B. conceived the concept. X.M. developed the methodology. X.M and O.V. collected and analyzed the data. X.M. and A.-L.B. were the lead writers of the manuscript.

Data Availability

The referenced dataset *unarXive* is available at https://doi.org/10.5281/zenodo.4313164. The referenced dataset of the full texts of *Nature* articles is provided by *Nature*. The textual dataset generated during the study is provided in the Supplementary Data. The code used for this manuscript is available at https://github.com/Barabasi-Lab/hidden-citation.

References

1 Garfield E. 1979. Is citation analysis a legitimate evaluation tool? *Scientometrics*. 1(4):359–375.
2 Evans JA. 2008. Electronic publication and the narrowing of science and scholarship. *Science*. 321(5887):395–399.
3 Uzzi B, Mukherjee S, Stringer M, Jones B. 2013. Atypical combinations and scientific impact. *Science*. 342(6157):468–472.
4 Merton RK. 1968. *Social theory and social structure*. enlarged ed. Boston: Free Press.
5 Garfield E. 1975. The obliteration phenomenon. *Curr Contents*. (51/52):5–7.
6 Einstein A. 1915. Die feldgleichungen der gravitation. In: Sitzungsberichte der Preussischen Akademie der Wissenschaften zu Berlin. p. 844–847.
7 Unruh WG. 1976. Notes on black-hole evaporation. *Phys Rev D*. 14(4):870–892.
8 Thomas KS. 1992. The development of eponymy; a case study of the southern blot. *Scientometrics*. 24(3):405–417.
9 Marx W, Cardona M. 2009. The citation impact outside references—formal versus informal citations. *Scientometrics*. 80(1):1–21.
10 Radicchi F, Fortunato S, Castellano C. 2008. Universality of citation distributions: toward an objective measure of scientific impact. *Proc Natl Acad Sci USA*. 105(45):17268–17272. https://doi.org/10.1073/pnas.0806977105.
11 Wang D, Song C, Barabási A-L. 2013. Quantifying long-term scientific impact. *Science*. 342(6154):127–132.
12 Zeng A, *et al.* 2017. The science of science: from the perspective of complex systems. *Phys Rep*. 714-715:1–73.
13 Fortunato S, *et al.* 2018. Science of science. *Science*. 359(6379): eaao0185.
14 Gerow A, Hu Y, Boyd-Graber J, Blei DM, Evans JA. 2018. Measuring discursive influence across scholarship. *Proc Natl Acad Sci USA*. 115(13):3308–3313.
15 Ma Y, Uzzi B. 2018. Scientific prize network predicts who pushes the boundaries of science. *Proc Natl Acad Sci USA*. 115(50): 12608–12615.
16 Wu L, Wang D, Evans JA. 2019. Large teams develop and small teams disrupt science and technology. *Nature*. 566(7744):378.
17 Zeng A, Fan Y, Di Z, Wang Y, Havlin S. 2021. Fresh teams are associated with original and multidisciplinary research. *Nat Hum Behav*. 5(10):1314–1322.
18 Peng H, Ke Q, Budak C, Romero DM, Ahn Y-Y. 2021. Neural embeddings of scholarly periodicals reveal complex disciplinary organizations. *Sci Adv*. 7(17):eabb9004.
19 McCain KW. 2011. Eponymy and obliteration by incorporation: the case of the "Nash equilibrium". *J Am Soc Inf Sci Technol*. 62(7):1412–1424.
20 McCain KW. 2012. Assessing obliteration by incorporation: issues and caveats. *J Am Soc Inf Sci Technol*. 63(11):2129–2139.
21 McCain KW. 2014. Assessing obliteration by incorporation in a full-text database: JSTOR, Economics, and the concept of "bounded rationality". *Scientometrics*. 101(2):1445–1459.
22 McCain KW. 2015. Mining full-text journal articles to assess obliteration by incorporation: Herbert A. Simon's concepts of bounded rationality and satisficing in economics, management, and psychology. *J Assoc Inf Sci Technol*. 66(11):2187–2201.
23 Cabanac G. 2014. Extracting and quantifying eponyms in full-text articles. *Scientometrics*. 98(3):1631–1645.
24 Schubert A, Glänzel W, Schubert G. 2022. Eponyms in science: famed or framed? *Scientometrics*. 127(3):1199–1207.
25 Cabanac G. 2018. What is the primordial reference for …?–Redux. *Scientometrics*. 114(2):481–488.

26 Maldacena J. 1999. The large-N limit of superconformal field theories and supergravity. *Int J Theor Phys*. 38(4):1113–1133.
27 Blei DM, Ng AY, Jordan MI. 2003. Latent Dirichlet allocation. *J Mach Learn Res*. 3(Jan):993–1022.
28 Jelodar H, *et al*. 2019. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimed Tools Appl*. 78(11):15169–15211.
29 Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. 2013. Distributed representations of words and phrases and their compositionality, https://doi.org/10.48550/arXiv.1310.4546, arXiv, arXiv:1310.4546, preprint: not peer reviewed.
30 Devlin J, Chang M-W, Lee K, Toutanova K. 2018. BERT: pre-training of deep bidirectional transformers for language understanding, https://doi.org/10.48550/arXiv.1810.04805, arXiv, arXiv:1810.04805, preprint: not peer reviewed.
31 Wang K, *et al*. 2020. Microsoft academic graph: when experts are not enough. *Quant Sci Stud*. 1(1):396–413.
32 King DA. 2004. The scientific impact of nations. *Nature*. 430(6997): 311–316.
33 Leahey E. 2006. Gender differences in productivity: research specialization as a missing link. *Gend Soc*. 20(6):754–780.
34 An W, Ding Y. 2018. The landscape of causal inference: perspective from citation network analysis. *Am Stat*. 72(3):265–277.
35 Huang J, Gates AJ, Sinatra R, Barabási A-L. 2020. Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proc Natl Acad Sci USA*. 117(9):4609–4616.
36 Small HG. 1978. Cited documents as concept symbols. *Soc Stud Sci*. 8(3):327–340.
37 Small H. 2004. On the shoulders of Robert Merton: towards a normative theory of citation. *Scientometrics*. 60(1):71–79.
38 Gates AJ, Ke Q, Varol O, Barabási A-L. 2019. Nature 's reach: narrow work has broad impact. *Nature*. 575(7781):32–34.
39 Merton RK. 1988. The Matthew effect in science, II: cumulative advantage and the symbolism of intellectual property. *ISIS*. 79(4):606–623.
40 Merton RK. 1968. The Matthew effect in science: the reward and communication systems of science are considered. *Science*. 159(3810):56–63.
41 Catalini C, Lacetera N, Oettl A. 2015. The incidence and role of negative citations in science. *Proc Natl Acad Sci*. 112(45): 13823–13826.
42 Buneman P, *et al*.. 2020. Why data citation isn't working, and what to do about it. *Database*. 2020(baaa022):1–16. https://doi.org/10.1093/databa/baaa022.
43 Sinha A, *et al*. 2015. An overview of Microsoft Academic Service (MAS) and applications. In: WWW '15 Companion: Proceedings of the 24th International Conference on World Wide Web. New York (NY): Association for Computing Machinery. p. 243–246. ISBN 978-1-4503-3473-0. https://doi.org/10.1145/2740908.2742839.
44 Saier T, Färber M. 2020. unarXive: a large scholarly data set with publications' full-text, annotated in-text citations, and links to metadata. *Scientometrics*. 125:3085–3108. https://doi.org/10.1007/s11192-020-03382-z.
45 Ginsparg P. 2009. The global village pioneers. *Learn Publ*. 22(2): 95–100.