RESEARCH ARTICLE

Frequently cocited publications: Features and kinetics

Sitaram Devarakonda[1], James R. Bradley[2], Dmitriy Korobskiy[1], Tandy Warnow[3], and George Chacko[1]

[1]Netelabs, NET ESolutions Corporation, McLean, VA, USA
[2]Raymond A. Mason College of Business, William and Mary, Williamsburg, VA, USA
[3]Department of Computer Science, University of Illinois Urbana-Champaign, Champaign, USA

Keywords: bibliometrics, cocitation

ABSTRACT

Cocitation measurements can reveal the extent to which a concept representing a novel combination of existing ideas evolves towards a specialty. The strength of cocitation is represented by its frequency, which accumulates over time. Of interest is whether underlying features associated with the strength of cocitation can be identified. We use the proximal citation network for a given pair of articles (*x*, *y*) to compute θ, an a priori estimate of the probability of cocitation between x and y, prior to their first cocitation. Thus, low values for θ reflect pairs of articles for which cocitation is presumed less likely. We observe that cocitation frequencies are a composite of power-law and lognormal distributions, and that very high cocitation frequencies are more likely to be composed of pairs with low values of θ, reflecting the impact of a novel combination of ideas. Furthermore, we note that the occurrence of a direct citation between two members of a cocited pair increases with cocitation frequency. Finally, we identify cases of frequently cocited publications that accumulate cocitations after an extended period of dormancy.

1. INTRODUCTION

Cocitation, "the frequency with which two documents from the earlier literature are cited together in the later literature," was first described in 1973 (Marshakova-Shaikevich, 1973; Small, 1973). As noted by Small (1973), cocitation patterns differ from bibliographic coupling patterns (Kessler, 1963) but align with patterns of direct citation and frequently cocited publications must have high individual citations.

Cocitation has been the subject of further study and characterization, for example, comparisons to bibliographic coupling and direct citation (Boyack & Klavans, 2010), the study of invisible colleges (Gmür, 2003; Noma, 1984), construction of networks by cocitation (Small & Sweeney, 1985; Small, Sweeney, & Greenlee, 1985), evaluation of clusters in combination with textual analysis (Braam, Moed, & van Raan, 1991), textual similarity at the article and other levels (Colavizza, Boyack, et al., 2018), and the fractal nature of publications aggregated by cocitations (van Raan, 1990).

Cocitations provide details of the relationship between key (highly cited) ideas, and changes in cocitation patterns over time may provide insight into the mechanism with which new schools of thought develop. Implicit in the definition of cocitation are novel combinations of existing ideas, but only some frequently cocited article pairs reflect surprising combinations. For example,

an open access journal

Citation: Devarakonda, S., Bradley, J. R., Korobskiy, D.,Warnow, T., & Chacko, G. (2020). Frequently cocited publications: Features and kinetics. *Quantitative Science Studies*, *1*(3), 1223–1241. https://doi.org/10.1162/qss_a_00075

DOI: https://doi.org/10.1162/qss_a_00075

Received: 27 March 2020
Accepted: 22 May 2020

Corresponding Author: George Chacko
george@nete.com

Handling Editor: Ludo Waltman

two publications presenting the leading methods for the same computational problem may be highly cocited, but this does not reflect a novel combination of ideas. Similarly, two publications describing methods that often constitute part of the same workflow may be highly cocited, but these cocitations are also not surprising. On the other hand, for two articles in different fields, frequent cocitation is generally unexpected.

Novel, atypical, or otherwise unusual combinations of cocited articles have been explored at the journal level (Boyack & Klavans, 2014; Bradley, Devarakonda, et al., 2020; Uzzi, Mukherjee, et al., 2013; Wang, Veugelers, & Stephan, 2017). However, journal-level classifications have limited resolution relative to article-level studies, which may better represent the actual structure and aggregations of the scientific literature (Gómez, Bordons, et al., 1996; Klavans & Boyack, 2017; Milojevic, 2019; Shu, Julien, et al., 2019; Waltman & van Eck, 2012). Accordingly, we sought to discover measurable characteristics of frequently cocited publications from an article-level perspective.

To study frequently cocited articles, we have developed a novel graph-theoretic approach that reflects the citation neighborhood of a given pair of articles. In seeking to determine the degree to which a cocited pair of papers represented a surprising combination, we wished to avoid journal-based field classifications, which present challenges. Instead, we attempted to use citation history to produce an estimate of the probability that a given pair of publications (x, y) would be cocited. As we focus on the activity before they are first cocited, the "probability" of cocitation is zero, by definition, because there are no cocitations yet. Hence, we approximated cocitation probabilities: We treat an article that cites one member of a cocited pair and also cites at least one article that cites the other member as a proxy for cocitation. Specifically, given a pair of publications x, y, we construct a directed bipartite graph whose vertex set contains all publications that cite either x or y previous to their first cocitation. We then compute θ, a normalized count of such proxies, and use it to predict the probability of cocitation between x and y. This approach enables an evaluation that is specific to the given pair of articles, and does so without substantial computational cost, while avoiding definitions of disciplines derived from journals or having to measure disciplinary distances.

To support our analysis, we constructed a data set of articles from Scopus (Elsevier BV, 2019) that were published in the 11-year period, 1985–1995, and extracted the cited references in these articles. Recognizing that frequently cocited publications must derive from highly cited publications (Small, 1973), we identified those reference pairs (33.6 million pairs) for each article in the data set that are drawn from the top 1% most cited articles in Scopus and measured their frequency of cocitation.

To investigate which statistical distributions might best describe the cocitation frequencies in these 33.6 million cocited pairs, we reviewed prior work on distributions of citation frequency (Eom & Fortunato, 2011; Newman, 2003; Price, 1965, 1976; Radicchi, Fortunato, & Castellano, 2008; Redner, 2005; Stringer, Sales-Pardo, & Amaral, 2008, 2010; Wang, Song, & Barabási, 2013). This research has fit the frequency distribution of citation strength sometimes to a power law distribution and other times to a lognormal distribution. A graph of the analogous cocitation data suggests that power law or lognormal distributions are candidates for describing cocitation strength as well and so we, accordingly, investigated that conjecture. Interestingly, Mitzenmacher (2003) notes that the debate between the appropriateness of power law versus lognormal distributions is not confined to bibliometrics, but has been at issue in many disciplines and contexts.

To study how the best-fit distributional function and parameters for cocitation might vary with θ, we stratified cocitation frequency data. We also measured whether a direct link exists between

two members of a cocited pair (i.e., whether one member of a pair cites the other) and how this property is related to cocitation frequencies. We find that the distribution of cocitation frequencies varies with θ and that a power law distribution fits cocitation frequencies more often when θ is small, whereas a lognormal distribution fits more often for large θ.

A pertinent aspect of cocitation is the rate at which frequencies accumulate. While the citation dynamics of individual publications have been fairly well studied by others, for example, Eom and Fortunato (2011) and Wallace, Larivière, and Gingras (2009), the dynamics of cocited articles are less well studied. Our interest was the special case analogous to the Sleeping Beauty phenomenon (Ke, Ferrara, et al., 2015; van Raan, 2004), which may reflect delayed recognition of scientific discovery and the causes attributed to it (Barber, 1961; Cole, 1970; Garfield, 1970, 1980; Glänzel & Garfield, 2004; Merton, 1963). Thus, we also identified cocited pairs that featured a period of dormancy before accumulating cocitations.

2. MATERIALS AND METHODS

2.1. Data

Citation counts were computed for all Scopus articles (88,639,980 records) updated through December 2019, as implemented in the ERNIE project (Korobskiy, Davey, et al., 2019). Records with corrupted or missing publication years or classified as "dummy" by the vendor were then removed, resulting in a data set of 76,572,284 publications. Hazen percentiles of citation counts, grouped by year of publication, were calculated for the these data (Bornmann, Leydesdorff, & Mutz, 2013). The top 1% of highly cited publications from each year were combined into a set of highly cited publications consisting of 768,993 publications.

Publications of type "article," each containing at least five cited references and published in the 11-year period from 1985–1995, were subsetted from Scopus to form a data set of 3,394,799 publications and 51,801,106 references (8,397,935 unique). For each of these publications, all possible reference pairs were generated and then restricted to those pairs where both members were in the set of highly cited publications (above).

For example, the data for 1985 consisted of 223,485 articles after processing as described above. Computing all reference pairs (that were also members of the highly cited publication set of 768,993) from these 223,485 articles gave rise to 2,600,101 reference pairs (Table 1) that ranged in cocitation frequency from 1 to 874 within the 1985 data set; from 1 to 11,949 across the 11-year period 1985–1995; and from 1 to 35,755 across all of Scopus. Collectively, the publications in our 1985–1995 data set generated 33,641,395 unique cocitation pairs, for which we computed cocitation frequencies across all of Scopus (Figure 1).

2.2. Derivation of θ

We now show how we define our prior on the probability of x and y being cocited, based on the citation graph restricted to publications that cite either x or y (but not both) up to the year of their first cocitation. Recall that we defined a proxy cocitation of x and y to be an article that cites one member of the cocited pair (x, y) and also cites at least one article that cites the other member. The idea behind this definition is that we consider papers that cite x as proxies for x, and papers that cite y as proxies for y. Thus, if a paper a cites both x and y' (where y' is a proxy for y), then it is a proxy for a cocitation of x and y. Similarly, if a paper b cites both y and x' (where x' is a proxy for x), it is also a proxy for a cocitation of x and y. This motivates the graph-theoretic formulation, which we now formally present.

Table 1. Summary of analyzed data. Publications of type "article" that had at least five cited references indexed in Scopus were selected from the 11 years 1985–1995. All possible reference pairs were generated for the cited references of these articles and then restricted to those pairs where both members were in the set of 768,993 highly cited publications. The fourth column shows the number of pairs in each year after the restriction was applied

Year	Articles	References	Cocited pairs
1985	223,485	1,796,502	2,600,101
1986	238,096	1,920,225	2,840,557
1987	250,575	2,037,654	3,180,261
1988	269,219	2,182,571	3,406,902
1989	285,873	2,303,481	3,793,986
1990	305,010	2,490,909	4,546,915
1991	325,782	2,662,005	5,039,334
1992	343,239	2,846,607	5,622,164
1993	360,916	3,006,374	6,121,147
1994	387,062	3,228,240	7,022,499
1995	405,503	3,432,228	7,626,684

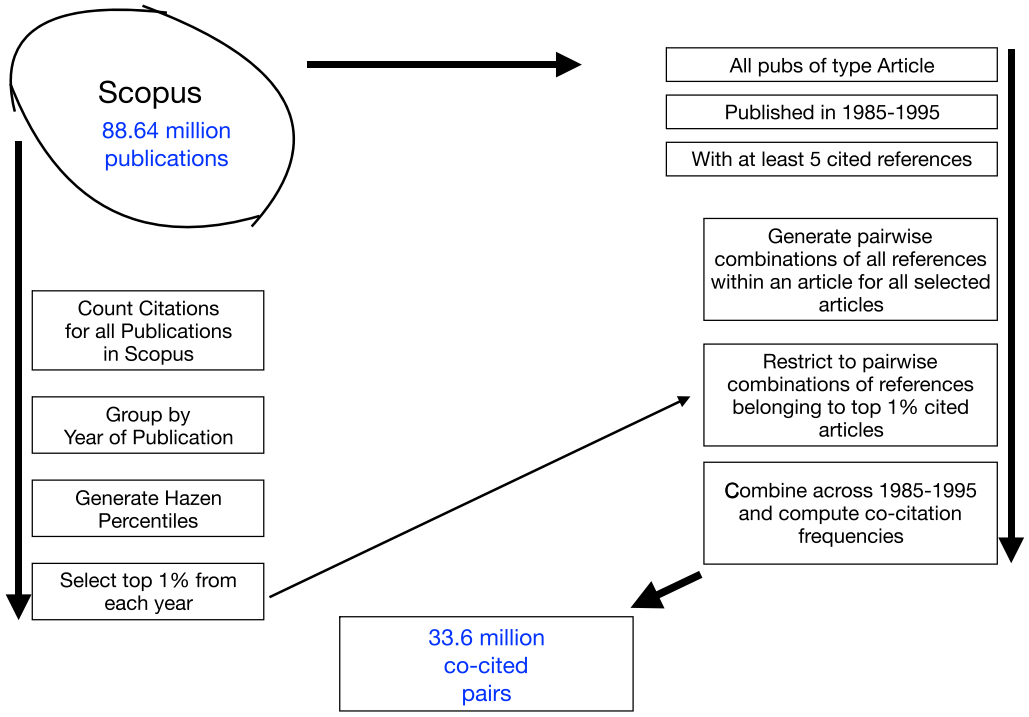


Figure 1. The workflow we used to generate a data set of 33,641,395 cocited publications from references cited by articles in Scopus published in the years 1985–1995.

We fix the pair x, y and we define $N(x)$ to be the set of all publications that cite x (but do not also cite y), and are published no later than the year of the first cocitation of x and y. We similarly define $N(y)$. We define a directed bipartite graph with vertex set $N(x) \cup N(y)$. Note that if x cites y then $x \in N(y)$, and similarly for the case where y cites x. Note also that because we have restricted $N(x)$ and $N(y)$ that $N(x) \cap N(y) = \emptyset$. We now describe how the directed edge set $E(x, y)$ is constructed. For any pair of articles a, b where $a \in N(x)$ and $b \in N(y)$, if a cites b then we include the directed edge $a \rightarrow b$ in $E(x, y)$. Similarly, we include edge $b \rightarrow a$ if b cites a. Finally, if a pair of articles both cite each other, then the graph has parallel edges. By construction, this graph is *bipartite*, which means that all the edges go between the two sets $N(x)$ and $N(y)$ (i.e., no edges exist between two vertices in $N(x)$, nor between two vertices in $N(y)$).

Note that by the definition, every edge in $E(x, y)$ arises because of a proxy cocitation, so that the number of proxy cocitations is the number of directed edges in $E(x, y)$. Consider the situation where a publication a cites x (so that $a \in N(x)$) and also cites b_1, b_2, b_3 in $N(y)$: this defines three directed edges from a to nodes of $N(y)$. We count this as three proxy cocitations, not as one proxy cocitation. Similarly, if we have a publication b that cites y and also cites a_1, a_2, a_3, a_4 in $N(x)$, then there are four directed edges that go from b to nodes in $N(x)$ and we will count each of those directed edges as a different proxy cocitation.

Accordingly, letting $|X|$ denote the cardinality of a set X, we note $|E(x, y)|$, (i.e., the number of directed edges that go between $N(x)$ and $N(y)$), is the number of proxy cocitations between x and y. If no parallel edges are permitted, the maximum number of possible proxy cocitations is $|N(x)| \times |N(y)|$. Under the assumption that both $N(x)$ and $N(y)$ each have at least one article, we define $\theta(x, y)$, our prior on the probability of x and y being cocited, as follows:

$$\theta(x, y) = \frac{|E(x, y)|}{|N(x)| \times |N(y)|}.$$

Note that if parallel edges do not occur in the graph, then $\theta(x, y) \leq 1$, but that otherwise the value can be greater than 1. Note also that $\theta(x, y) = 0$ if $E(x, y) = \emptyset$ (i.e., if there are no proxy cocitations) and that $\theta(x, y) = 1$ if every possible proxy cocitation occurs.

To efficiently calculate θ, we used the following pipeline. We copied Scopus data from a relational schema in PostgreSQL into a citation graph from Scopus into the Neo4j 3.5 graph database using an automated Extract Transform Load (ETL) pipeline that combined Postgres CSV export and the Neo4j Bulk Import tool. The graph vertex set is all publications, each with a publication year attribute, and the edge set is all citations between the publications. A Cypher index was created on the publication year. We developed Cypher queries to calculate θ and tuned performance by splitting input publication pairs into small batches and processing them in parallel, using parallelization in Bash and GNU Parallel. Batch size, the number of parallel job slots, and other parameters were tuned for performance, with best results achieved on batch sizes varying from 20 to 100 pairs. The results of θ calculations were cross-checked using SQL calculations. In the small number of cases where θ computed to >1 (above) it was set to 1 for the purpose of this study.

2.3. Statistical Calculations

We denote the observed cocitation frequency data by the multiset

$$X^o = \{x_1^o, \ldots, x_N^o\},$$

where N is the total number of pairs of articles and x_i^o is the observed frequency of the ith pair of papers being cocited. Note that this is in general a multiset, as different pairs of articles can have the same cocitation frequency. Let $n(x)$ be the number of times that x appears in X^o (equivalently,

$n(x)$ is the number of pairs of articles that are cocited x times), and let $N(x) = \sum_{y=x}^{\infty} n(y)$ denote the total number of pairs of articles that are cocited at least x times. Then

$$f^o(x \mid x \geq \underline{x}) = \frac{n(x)}{N(\underline{x})} \text{ for } x \in [\underline{x}, \infty), \tag{1}$$

where $\underline{x}$ is a parameter we use to analyze the distribution's right tail starting at varying frequencies. We describe in this subsection (a) the statistical computations for fitting log-normal and power law distributions to right tails of the observed cocitation frequency distributions as defined by Eq. 1 for various $\underline{x}$ and (b) how we assessed the quality of those fits. Further, we performed such analyses for various slices of the data, stratifying by θ and other parameters, as is described in Section 3.

We used a discrete version of a lognormal distribution to represent integer cocitation frequencies, $f(\cdot)$, following Stringer et al. (2008) and Stringer et al. (2010), while appropriately normalizing for our conditional assessment of the right tail commencing at $\underline{x}$:

$$f_{LN}(x \mid \mu, \sigma, \underline{x}) = \frac{\widetilde{f}(x \mid \mu, \sigma)}{\sum_{n=\underline{x}}^{\infty} \widetilde{f}(n \mid \mu, \sigma)} \text{ for } x \geq \underline{x} \tag{2}$$

$$\widetilde{f}(x \mid \mu, \sigma) = \int_{x-0.5}^{x+0.5} \frac{dq}{q\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln q - \mu)^2}{2\sigma^2}\right),$$

where μ and σ are the mean and standard deviation, respectively, of the underlying normal distribution. These probabilities can be computed with the cumulative normal distribution,

$$\widetilde{f}(x \mid \mu, \sigma) = \Phi\left(\frac{\ln(x+0.5)}{\sigma}\right) - \Phi\left(\frac{\ln(x-0.5)}{\sigma}\right),$$

using the well-known error function.

We fit distributions to the cocitation frequency data for various extremities of the right tail, as parameterized by $\underline{x}$, using a maximum (log) likelihood estimator (MLE). We solved for the best-fit distributional parameters for the lognormal distribution, μ and σ, by modifying a multidimensional interval search algorithm from Press, Teukolsky, et al. (2007) and following Stringer et al. (2010). A compiled version of this code using the C++ header file `amoeba.h` is available on our Github site (Korobskiy et al., 2019).

We fit a discrete power law distribution to the data for various values of $\underline{x}$, which was normalized for our conditional observations of the right tail:

$$f_{PL}(x \mid \alpha, \underline{x}) = \frac{x^{-\alpha}}{\zeta(\alpha, \underline{x})} \text{ for } x \geq \underline{x}, \tag{3}$$

where the Hurwitz zeta function,

$$\zeta(\alpha, \underline{x}) = \sum_{x=0}^{\infty} \frac{1}{(x + \underline{x})^{\alpha}},$$

is a generalization of the Riemann zeta function, $\zeta(\alpha, 1)$, as is needed for analysis of the right tail.

We solved first-order conditions for the (log) MLE to find the best-fit distributional exponent α,

$$\frac{\zeta'(\alpha, \underline{x})}{\zeta(\alpha, \underline{x})} = -\frac{1}{N(\underline{x})} \sum_{x \in X^o(\underline{x})} \ln x, \tag{4}$$

as described in Clauset, Shalizi, and Newman (2009) and Goldstein, Morris, and Yen (2004), where $X^o(\underline{x}) = \{x \in X^o : x \geq \underline{x}\}$, are the observed cocitations with frequencies at least as great as $\underline{x}$ and $N(\underline{x})$ is the number of such cocitations. We solved Eq. 4 to find α using a bisection algorithm.

We used the χ^2 goodness of fit (χ^2) and the Kolmogorov-Smirnov (K-S) tests to assess the null hypothesis that the distribution of the observed cocitation frequencies and the best-fit lognormal distribution are the same, and similarly for the best-fit power law distribution. We also computed the Kullback-Leibler Divergence (K-L) between the observed data and the best-fit distributions.

Both the χ^2 and K-S tests employed the null hypothesis that the observed cocitation frequencies, $n(x)$ for $x \in [\underline{x}, \infty)$, were sampled from the best-fit lognormal or power law distributions, which we denote by $f_d(\cdot|\underline{x})$ for $d \in \{LN, PL\}$, while suppressing the parameters specific to each of the distributions.

The usual χ^2 statistic was computed by, first, grouping each of the observed cocitation frequencies into k bins, denoted by b_i for $i \in \{1, \ldots, k\}$, and then computing

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

where O_i is the observed number of cocitations having frequencies associated with the ith bin,

$$O_i = \sum_{x \in b_i} n(x),$$

and E_i is the expected number of observations for frequencies in bin i, if the null hypothesis was true, in a sample with size equal to the number of observed data points, $N(\underline{x})$:

$$E_i = \sum_{x \in b_i} f_d(x|\underline{x})\ N(\underline{x})$$

If the null hypothesis is true, then we would expect O_i and E_i to be approximately equal, with deviations owing to variability due to sampling.

Constructing the bins b_i requires only that $E_i \geq 5$ for every $i = 1, \ldots, k$. Test outcomes are sometimes sensitive to the minimum E_i permitted, which we will denote by $\underline{E}$, so we tested with multiple thresholds, including 10, 20, 50, and 70. Furthermore, statistical tests are stochastic: These multiple tests permitted a reduction in the probability of erroneously rejecting or accepting the null hypothesis based on a single test. The distribution of observed cocitation frequencies was skewed right with a long tail, so that aggregating bins to satisfy $E_i \geq \underline{E}$ was most critical in the right tail. This motivated a bin construction algorithm that aggregated frequencies in reverse order, starting with the extreme right tail. Algorithm 1 requires a set of the unique observed cocitation frequencies, $\hat{X}^o$, which includes the elements of the multiset X^o without repetition. While Algorithm 1 does not guarantee in general that all bins satisfy $E_i \geq \underline{E}$, that criterion was satisfied for the observed data.

We implemented a K-S test using simulation to generate a sampling distribution to account for the discrete frequency observations (StackExchange, 2014). We denote the cumulative distribution of observed cocitation frequencies by $F^o(x|\underline{x}) = \sum_{i=\underline{x}}^{x} f^o(i|\underline{x})$, and the best-fit cumulative

Algorithm 1: Frequency bin construction

```
1: i ← 1
2: b_1 = {}
3: while |X̂^o| > 0 do
4:     b_i ← b_i ∪ {max (X̂^o)}
5:     X̂^o ← X̂^o \ max (X̂^o)
6:     if E_i ≥ E̲ then
7:         i ← i + 1
8:         b_i ← {}
9:     end if
10: end while
```

distribution by $F_d(x|\underline{x}) = \sum_{i=\underline{x}}^{x} f_d(i|\underline{x})$. The K-S test involves testing the maximum absolute difference between the observed and theorized cumulative distributions,

$$D_n = \max_x \left| F^o\left(x|\underline{x}\right) - F_d\left(x|\underline{x}\right) \right|,$$

where n is the number of observations giving rise to $F^o(x|\underline{x})$, against the distribution of such differences between samples from the theorized distribution with the same number of observations, n,

$$\widetilde{D}_n = \max_x \left| \widetilde{F}_{d,1}\left(x|\underline{x}\right) - \widetilde{F}_{d,2}\left(x|\underline{x}\right) \right|,$$

where $\widetilde{F}_{d,j}(x|\underline{x})$ is the empirical distribution of sample j of size n (notation suppressed) drawn from $F_d(x|\underline{x})$. We generated 100 such random variables $\widetilde{D}_n$ for each test. We reject the null hypothesis if D_n is larger than substantially all of the $\widetilde{D}_n$, say all but 5%, for equivalence with a p-value of 0.05. The number of $\widetilde{D}_n$ samples drawn yields a p-value with a resolution of 1%.

We computed the K-L Divergence two ways due to its asymmetry:

$$D_{K\text{-}L}(f^o \| f_d) = \sum_{x=\underline{x}}^{\infty} f^o\left(x|\underline{x}\right) \ln \frac{f^o\left(x|\underline{x}\right)}{f_d\left(x|\underline{x}\right)}$$

$$D_{K\text{-}L}(f_d \| f^o) = \sum_{x=\underline{x}}^{\infty} f_d\left(x|\underline{x}\right) \ln \frac{f_d\left(x|\underline{x}\right)}{f^o\left(x|\underline{x}\right)}.$$

Separate from the tests above, we tested whether the distribution of cocitation frequencies was independent of θ using a χ^2 test, using the null hypothesis that the cocitation frequency distribution was independent of θ. We initially created a contingency table on θ and cocitation frequency using these bins for θ, {[0.0, 0.2), [0.2, 0.4), [0.4, 0.6), [0.6, 0.8), [0.8, 1.0]}, and logarithmic bins for frequency to accommodate the skewed distributions:

$$\{[10, 100), [100, 1000), [1000, 10000), [10000, 100000]\}.$$

We subsequently aggregated these bins to have an expected number of cocitations in each bin equal to or greater than 5 to account for a decreasing number of observations as θ and frequency increased by having just two intervals for frequency: {[10, 100), [100, 100000]}.

2.4. Kinetics of Cocitation

We extended prior work on delayed recognition and the Sleeping Beauty phenomemon (Glänzel & Garfield, 2004; Ke et al., 2015; Li & Ye, 2016; van Raan, 2004) towards cocitation. We have modified the beauty coefficient (B) of Ke et al. (2015) to address cocitations by (a) counting citations to a pair of publications (cocitations) rather than citations to individual papers, (b) setting t_0 (age zero) to the first year in which a pair of publications could be cocited (i.e., the publication year of the more recently published member of a cocited pair), and (c) setting C_0 to the number of cocitations occurring in year t_0. Rather than calculate awakening time as in Ke et al. (2015), we opted to measure the simpler length of time between t_0 and the first year in which a cocitation was recorded; we label this measurement as the time lag t_l, so that $t_l = 0$ if a cocitation was recorded in t_0.

3. RESULTS AND DISCUSSION

Our base data set, described in Table 1, consists of the 33,641,395 cocited reference pairs (33.6 million pairs) and their cocitation frequencies, gathered from Scopus during the 11-year period from 1985–1995 (Section 2). A striking distribution of cocitation frequencies with a long right tail is observed with a minimum cocitation of 1, a median of 2, and a maximum cocitation frequency of 51,567 (Figure 2). Approximately 33.3 of 33.6 million pairs (99% of observations) have cocitation frequencies ranging from 1–67 and the remaining 1% have cocitation frequencies ranging from 68–51,567. As the focus of our study was cocitations of frequently cited publications, we further restricted this data set to those pairs with a cocitation frequency of at least 10, which resulted in a smaller data set of 4,119,324 cocited pairs (4.1 million pairs) with minimum cocitation frequency of 10, median of 18, and a maximum cocitation frequency of 51,567. To focus on cocitations derived from highly cited publications, θ was calculated for all pairs with a cocitation frequency of at least 10. We also note whether one article in a cocitation pair

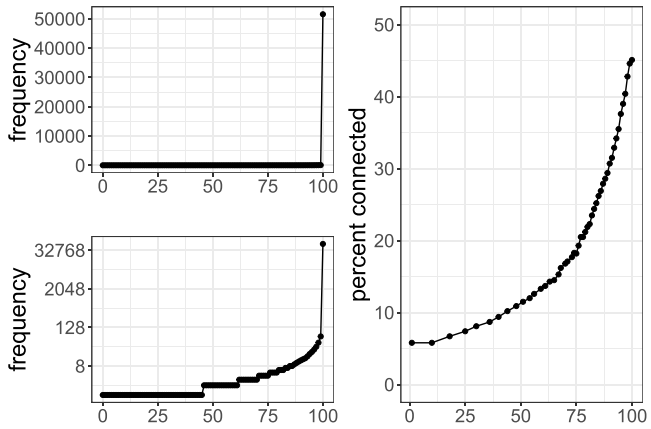


Figure 2. The *x*-axis shows percentiles for all three plots. Left: Cocitation frequencies of highly cited publications from Scopus 1985–1995. Cocitation frequencies are plotted against their percentile values. The upper and lower plots were both generated from 33,641,395 data points. The lower plot shows the same data with a logarithmic (ln) transformation of the *y*-axis. The minimum cocitation frequency is 1, the median is 2, the third quartile is 4, and the maximum is 51,567. Additionally; 15,140,356 pairs (45%) have a cocitation frequency of 1. Frequencies of 12, 22, 67, and 209 correspond to quantile values of 0.9, 0.95, 0.99, and 0.999 respectively. Right: Direct citations between members of a cocited pair (connectedness) increase with cocitation frequency. The proportion of connected pairs (a direct citation exists between the two members of a pair) within each percentile is shown. Data are plotted for all pairs with a cocitation frequency of at least 10 (4.1 million pairs).

cites the other (connectedness), reporting a pair as "connected" when such a citation occurs, else as "not connected."

Influenced by the use of linked cocitations for clustering (Small & Sweeney, 1985), we also examined the extent to which members of a cocited pair were also found in other cocited pairs. We found that 205,543 articles contributed to 4.12 million cocited pairs. The highest frequency observed in our data set, 51,567 cocitations, was for a pair of articles from the field of physical chemistry: Becke (1993) and Lee, Yang, and Parr (1988). The members of this pair are not connected and are found in 1,504 cocited pairs with frequencies ranging from 10 to 51,567. The second highest frequency, 28,407 cocitations, was for another pair of articles from the field of biochemistry: Bradford (1976) and Laemmli (1970). Members of this pair are not connected and are found in a staggering 41,909 cocited pairs, 24,558 for the Laemmli gel electrophoresis article and 17,352 for the Bradford protein estimation article. For the latter pair, both articles describe methods heavily used in biochemistry and molecular biology, an area with strong referencing activity, so this result is not entirely surprising.

Having developed $\theta(x, y)$ as a prediction of the probability that articles x and y would be cocited, we first tested whether the distribution of cocitation frequencies was independent of θ (Section 2). The null hypothesis that the cocitation frequency distribution was independent of θ was rejected with a very small p-value: The statistical software indicated a p-value with no significant nonzero digits. We next investigated what distribution functions might fit the frequencies of cocitation as θ varied.

Based on the long tails of citation frequencies, prior research has assessed the fit of log-normal and power law distributions (Radicchi et al., 2008; Stringer et al., 2008, 2010). We noted long right tails in cocitation frequencies, which, similarly, motivated us to assess the fit of lognormal and power law distributions to cocitation data. Further, we stratified the data according to (a) the minimum frequency for the right tail $\underline{x}$, (b) θ, and (c) whether the two members of each cocitation pair were connected. Figure 3 shows which distribution, if either, fits the data in each slice, based on tests of statistical significance. Note that there were no circumstances where both distributions fit: If one fit, then the other did not.

Statistical tests were not possible for some slices due to an insufficient number of data points. This was the case for certain combinations of large $\underline{x}$, large θ, and cocitations that were not connected. The number of data points obviously decreases as $\underline{x}$ increases, and we found the decrease in the number of data points to be more precipitous when θ was large and cocitations were unconnected due to the lighter right tails for these parameter combinations. The graph in the right panel of Figure 4, which has a logarithmic y-axis, shows that the number of data points per θ interval analyzed decreases most often by more than an order of magnitude from one interval to the next as θ increases. Most pairs of publications that are cocited at least 10 times, therefore, have small values of θ.

Figure 3 indicates when the null hypothesis of a best-fit lognormal or power law fitting the observed data cannot be rejected. We computed two types of statistics for evaluating the null hypothesis (χ^2 and K-S) and, moreover, we computed the χ^2 statistic for four binning strategies. Figure 3 indicates a distributional fit, specifically, if either the K-S p-value is greater than 0.05 or if two or more of the χ^2 statistics are greater than 0.05. While we computed the K-L Divergence (see supplementary material), we did not use these computations for formal statements of distributional fit because they are neither a norm nor determine statistical significance. These K-L computations did, however, support the findings based on formal tests of statistical significance.

Power law distributions fit most often when cocitations are connected (Figure 3), when more extreme right tails are considered, and when cocitations have small values of θ. Log-normal

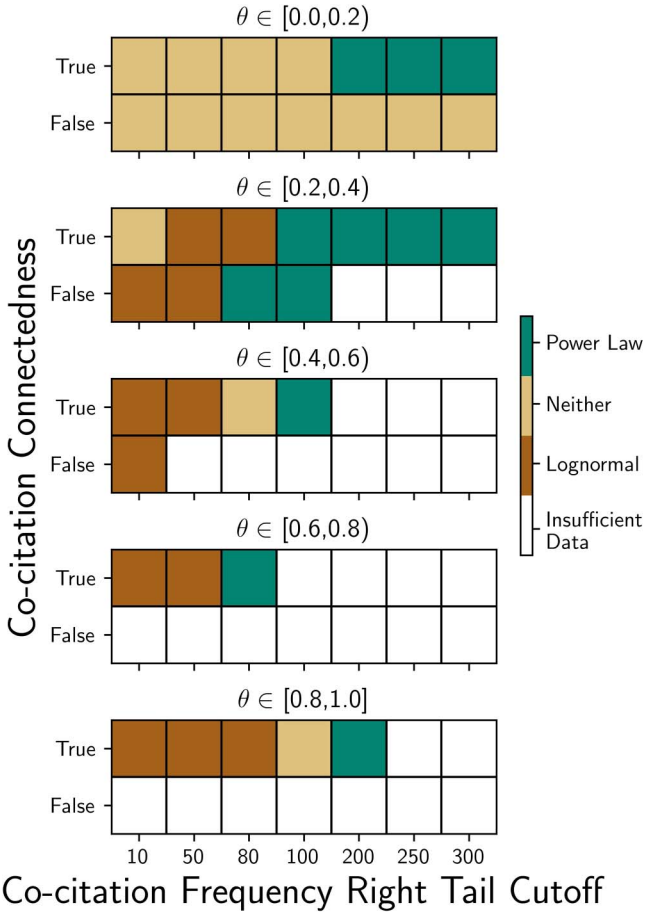
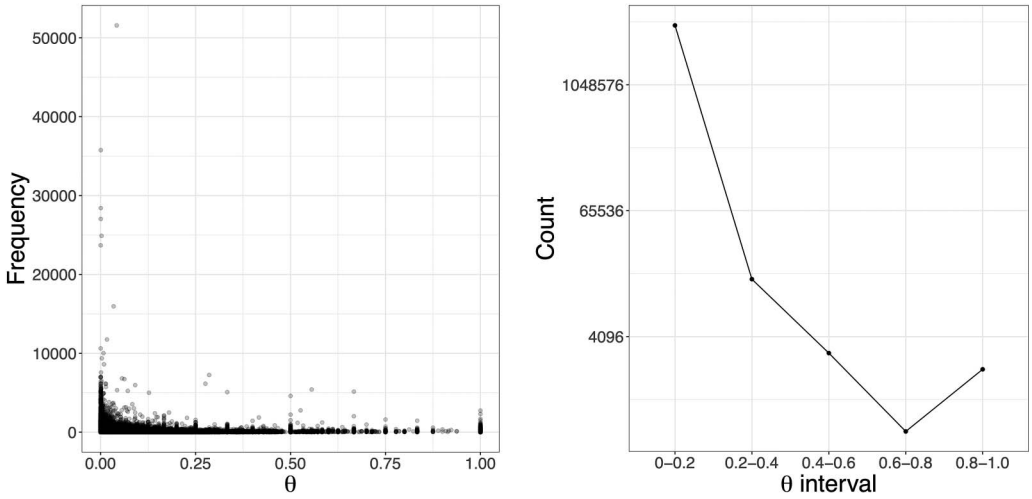


Figure 3. Distributional fits to the observed cocitation frequencies. The graph shows where a lognormal or power law distribution demonstrated a statistically significant fit with the observed cocitation frequencies stratified by θ, extent of the right tail tested $\underline{x}$, and whether cocitations were connected. A power law fit more often fore in the intervals [0.0, 0.2) and [0.2, 0.4) when cocitation constituents were connected. When a lognormal distribution fit, it was for broader portions of the data set. Data were insufficient for testing as θ increased due to (a) fewer observations and (b) less prominent right tails.

distributions fit, conversely, in some circumstances, when a greater portion of the right tail is considered. These observations support the existence of heavy tails for θ small, even if a lognormal distribution fits the observed data more broadly. This observation is consistent with our observations of the most frequent cocitations having small θ values, as shown in the scatter plot in the left panel of Figure 4.

Mitzenmacher (2003) shows a close relationship between the power law and lognormal distributions vis-à-vis subtle variations in generative mechanisms that determine whether the resulting distribution is a power law or lognormal. The stratified layers in Figure 3, where a lognormal distribution fits for some portion of the right tail and, in the same instance, a power law describes the more extreme tail, may, therefore, be due to a generative mechanism whose parameters are close to those for a power law distribution as well as those for a lognormal distribution.

Table 2 shows the exponents of the best-fit power law distributions when statistical tests indicated that a power law was a good fit and where comparisons were possible among the intervals



(a) Co-citation Scopus frequency versus θ (b) Number of co-cited pairs per θ interval

Figure 4. Cocitation dynamics relative to θ. (a) Points represent the Scopus frequency vs. θ value for each cocited pair. Darker regions indicate denser plots of the translucent points. cocited pairs with the greater frequency are observed for pairs with smaller θ. (b) The y-axis employs a log scale and shows the number of cocited pairs per θ interval. The number of cocited pairs decreases, most often, by more than an order of magnitude per interval as θ increases. The dominance of cocited pairs with smaller θ are also reflected by regions of greater density in panel (a).

of θ: These were possible for θ intervals of [0.0, 0.2) and [0.2, 0.4), for connected cocitations, and right tails commencing at $\underline{x} \in \{200, 250, 300\}$. The power law exponent α in these comparisons was less for $\theta \in [0.0, 0.2)$ than for $\theta \in [0.2, 0.4)$, indicating heavier tails for θ small and, therefore, a greater chance of extreme cocitation frequency. Figure 5 shows a log-log plot of the number of cocitations (y-axis) exhibiting the counts on the x-axis, for θ in the interval [0.0, 0.2) (note that both axes employ log scaling). The pattern for points below the 99th percentile clearly indicates that the number of cocitations referenced at a given frequency decreases greatly as the frequency increases. Also, the broadening of the scatter where fewer cocitations are cited more frequently is indicative of a long right tail, as has been observed in other research where lognormal or power law distributions have been fit to data, as in Montebruno, Bennett, et al. (2019).

Table 2. Exponents of best-fit power law distributions. These observations are for power law exponents where comparison across intervals of θ were possible, and where statistical tests indicated that a power law was a good fit to the data. The articles of the cocitations were connected for all data shown

Right tail cutoff ($\underline{x}$)	θ	Power law exponent (α)
200	[0.0, 0.2)	3.26
200	[0.2, 0.4)	3.37
250	[0.0, 0.2)	3.27
250	[0.2, 0.4)	3.37
300	[0.0, 0.2)	3.22
300	[0.2, 0.4)	3.35

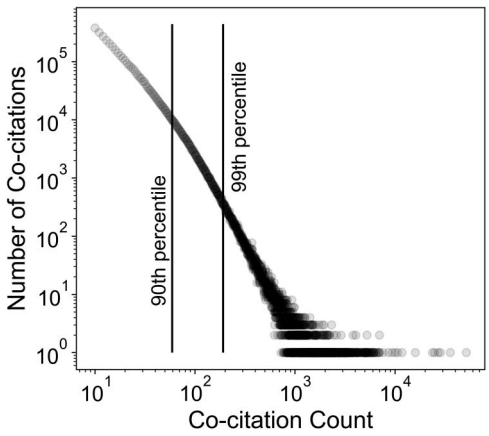


Figure 5. Log-log plot of the number of cocitations versus cocitation count for $\theta \in [0.0, 0.2)$. The *y*-axis shows the number of cocited pairs observed having the citation counts plotted along the *x*-axis. The tightly clustered plot below the 99th percentile demonstrates a clear pattern of decreasing number of cocited pairs having an increasing number of citation counts. The scatter plot for the tail above the 99th percentile broadens, indicating a long tail of relatively few cocited pairs that were cited with extreme frequency.

Perline (2005) warns against fitting a power law function to truncated data. Informally, a portion of the entire data set can appear linear on a log-log plot, while the entire data set would not. He cites instances where researchers have mistakenly characterized an entire data set as following a power law due to an analysis of only a portion of the data, when a lognormal distribution might provide a better fit to the entire data set. Indeed, the scatter plot in Figure 5 is not linear and so, as Figure 3 shows, a power law does not fit the entire data set. This is what Perline calls a weak power law, where a power law distribution function fits the tail but not the entire distribution. Our concern, however, is not with characterizing the distributional function for the entire data set, but with characterizing the features of high frequency cocitations, which by definition means we are concerned with the right tail of the distribution. Moreover, the results avoid confusion between lognormal and power law distribution functions because we have shown not only that a power law provides a statistically significant fit but also that a lognormal distribution function does not fit.

Our analysis found particularly heavy tails that were well fit by power law distributions for small θ, in the intervals [0.0, 0.2) and [0.2, 0.4), and for cocitations whose constituents are connected, as shown in Figure 3. The closely related Matthew Effect (Merton, 1968), cumulative advantage (Price, 1976), and the preferential attachment class of models (Albert & Barabási, 2002) provide possible explanations for citation frequencies following a power law distribution for some sufficiently extreme portion of the right tail. For greater values of θ, insufficient data in the right tails precludes a definitive assessment in this regard, although one might argue that the lack of observations in the tails is counter to the existence of a power law relationship. It is also noteworthy that the exponents we found for cocitations (Table 2) are close in value to those reported for citations by Price (1976) and Radicchi et al. (2008).

3.1. Delayed Cocitations

The delayed onset of citations to a well-cited publication, also referred to as *Delayed Recognition* and *Sleeping Beauty*, has been studied by Garfield, van Raan, and others (Bornmann, Ye, & Ye, 2018; Garfield, 1970; Glänzel & Garfield, 2004; Ke et al., 2015; Li & Ye, 2016; van Raan, 2004; van Raan & Winnink, 2019). We sought to extend this concept to frequently cocited articles

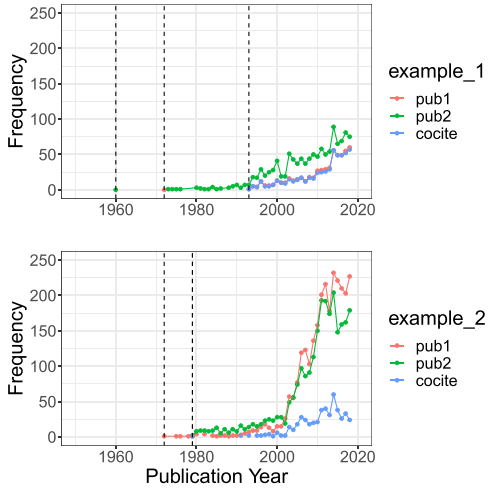


Figure 6. Cocitation frequencies of highly cited publications from Scopus 1985–1995. *Upper panel*: Publication 1: Instability of the interface of two gases accelerated by a shock wave (1972) https://doi.org/10.1007/BF01015969, first cited (1993), total citations (566). Publication 2: Taylor instability in shock acceleration of compressible fluids (1960) https://doi.org/10.1002/cpa.3160130207, first cited (1973), total citations (566), first cocited (1993), total cocitations (541). *Lower Panel*: Publication 1: Colorimetric assay of catalase (1972) https://doi.org/10.1016/0003-2697(72)90132-7, first cited (1972), total citations (2,683). Publication 2: Levels of glutathione, glutathione reductase and glutathione S-transferase activities in rat lung and liver (1979) https://doi.org/10.1016/0304-4165(79)90289-7, first cited (1979), total citations (2,464), first cocited (1979), total cocitations (470).

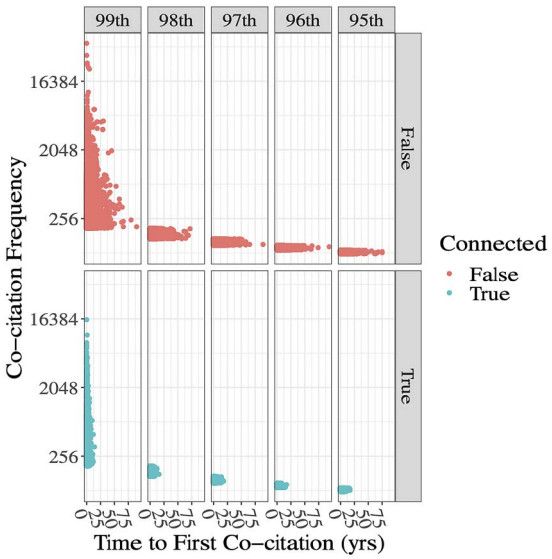


Figure 7. Relationship between time lag (t_l) and cocitation frequency. Extended lag times are associated with lower cocitation frequencies. Connected pairs have lower t_l values. Data are shown for 207,214 pairs consisting of ≥ 95th percentile of cocitation frequencies for the 4.1 million row data set. The observations are stratified by percentile group (vertical panels) and connectedness (upper and lower halves). Cocitation frequency (y-axis) is plotted against t_l, the time between first possible cocitation and first cocitation.

(Figure 6). As an initial step, we calculated two parameters (Section 2): (a) the beauty coefficient (Ke et al., 2015) modified for cocited articles and (b) timelag t_l, the length of time between first possible year of cocitation and the first year in which a cocitation was recorded. We further focused our consideration of delayed cocitations to the 95th percentile or greater of cocitation frequencies in our data set of 4.1 million cocited pairs. Within the bounds of this restriction, 24 cocited pairs have a beauty coefficient of 1,000 or greater and all 24 are in the 99th percentile of cocitation frequencies. Thus, very high beauty coefficients are associated with high cocitation frequencies.

We also examined the relationship of t_l with cocitation frequencies (Figure 7) and observed that high t_l values were associated with lower cocitation frequencies. These data appear to be consistent with a report from van Raan and Winnink (van Raan & Winnink, 2019), who conclude that "probability of awakening after a period of deep sleep is becoming rapidly smaller for longer sleeping periods." Further, when two articles are connected, they tend to have smaller t_l values compared to pairs that are not connected in the same frequency range.

4. CONCLUSIONS

In this article, we report on our exploration of features that impact the frequency of cocitations. In particular, we wished to examine article pairs with high cocitation frequencies with respect to whether they originated from the same school(s) of thought or represented novel combinations of existing ideas. However, defining a discipline is challenging, and determining the discipline(s) relevant to specific publications remains a challenging problem. Journal-level classifications of disciplines have known limitations and while article-level approaches offer some advantages, they are not free from their own limitations (Milojevic, 2019).

Consequently, we designed θ, a statistic that examines the citation neighborhood of a pair of articles x and y to estimate the probability that they would be cocited. Our approach has advantages compared to alternate approaches: It avoids the challenges of journal-level analyses, it does not require a definition of "discipline" (or "disciplinary distance"), it does not require assignment of disciplines to articles, it is computationally feasible, and, most importantly, it enables an evaluation that is specific to a given pair of articles.

We note that when x and y are from the same subfield, then θ may be very large, and conversely, when x and y are from very different fields, it might be reasonable to expect that θ will be small. Thus, in a sense, θ may correlate with disciplinary similarity, with large values for θ reflecting conditions where the two publications are in the same (or very close) subdisciplines, and small values for θ reflecting that the disciplines for the two publications are very distantly related. We also comment that in this initial study, we have not considered second-degree information, that is, publications that cite publications that cite an article of interest.

Our data indicate that the most frequent cocitations occur when cocitations have small values of θ, as shown in Figure 4. Our study considered the hypothesis that the frequency distribution is independent of θ, but our statistical tests rejected this hypothesis, and showed instead that the frequency distribution is best characterized by a power law for small values of θ and connected publications, and in many other regions is best characterized by a lognormal distribution.

The observation that power laws are consistent with small values of θ and connected cocitations is consistent with the theory of preferential attachment for these parameter settings. To the extent that preferential attachment is the mechanism giving rise to a power law, this suggests that preferential attachment is, at least, stronger for small θ values and connected cocitations than for other parameter combinations, or that preferential attachment is not applicable to other parameter values.

Observing power laws, heavy tails, and pairs with extreme cocitation strength for small values of θ (i.e., pairs that have small a priori probabilities of being cocited) may seem, on its face, paradoxical. One possible explanation for the pairs in the extreme right tail with both small θ and large cocitation strength is that those pairs represent novel combinations of ideas that, when recognized within the research community, catalyze an increased citation rate, consistent with preferential attachment coupled to time-dependent initial attractiveness (Eom & Fortunato, 2011) as an underlying generative mechanism. However, small values of θ do not guarantee a high cocitation count: Indeed, even for small values of θ, cocitations with a power law predominantly have relatively low cocitation strength.

We also note the increasing proportion of connected pairs as the percentile for cocitation frequency increases (Figure 2); this pair of parameters appears to be associated with a fertile environment where extremely high cocitation frequencies are possible. This observation raises the question of whether small values of θ and connected cocitations are associated with preferential attachment and, if a causal relationship exists, then how do θ and cocitation connection provide an environment supporting preferential attachment? A possibility is that one article in a cocited pair citing the other makes the potential significance of the combination of their ideas apparent to researchers. The clear pattern of the highest frequency cocited pairs typically having low θ values suggests that these pairs are highly cited and hence impactful because of the novelty in the ideas or fields that are combined (as reflected in low θ). However, other factors should be considered, such as the prominence of authors and prestige of a journal (Garfield, 1980) where the first cocitation appears.

We did not apply field normalization techniques when assembling the parent pool of 768,993 highly cited articles consisting of the top 1% of highly cited articles from each year in the Scopus bibliography. Thus, the highly cocited pairs we observe are biased towards high-referencing areas such as biomedicine and parts of the physical sciences (Small & Greenlee, 1980). However, the data set we analyzed has a lower bound of 10 on cocitation frequencies and includes pairs from fields other than those that are high referencing. For example, the maximum t_l we observed in the data set of 4.1 million pairs was 149 years, and is associated with a pair of articles independently published in 1840, establishing their eponymous Staudt-Clausen theorem (Clausen, 1840; von Staudt, 1840); this pair of articles has apparently been cocited 10 times since their publication. A second pair of articles concerning electron theory of metals (Drude, 1900a, 1900b),was first cocited in 1994, 109 times, with t_l observed of 94 years. Both cases are drawn from mathematics and physics rather than the medical literature. They are also consistent with the suggestion that the probability of awakening is smaller after a period of deep sleep (van Raan & Winnink, 2019). As we have defined t_l, with its heavy penalty for early citation, we create additional sensitivity to coverage and data quality especially for pairs with low citation numbers. Indeed, for the Staudt-Clausen pair, a manual search of other sources revealed an article (Carlitz, 1961) in which they are cocited. Both these articles were originally published in German and it is possible that additional cocitations were not captured. Thus, big data approaches that serve to identify trends should be accompanied by more meticulous case studies, where possible. Other approaches for examining depth of sleep and awakening time should certainly be considered (Ke et al., 2015; van Raan, 2004; van Raan & Winnink, 2019). Lastly, using our approach to revisit invisible colleges (Crane, 1972; Price & Beaver, 1966; Small & Sweeney, 1985) seems warranted, as it seems likely that the upper bound of a hundred members predicted by Price and Beaver (1966) is likely to have increased in a global scientific enterprise with electronic publishing and social media.

Finally, we view these results as a first step towards further investigation of cocitation behavior, and we introduce a new technique based on exploring first-degree neighbors of cocited

publications; we are hopeful that this graph-theoretic study will stimulate new approaches that will provide additional insights, and prove complementary to other article-level approaches.

ACKNOWLEDGMENTS

We thank two anonymous reviewers for their helpful and constructive critique. In addition to support through federal funding, the ERNIE project features a collaboration with Elsevier. We thank our colleagues from Elsevier for their support of the collaboration.

AUTHOR CONTRIBUTIONS

Sitaram Devarakonda: Conceptualization, Methodology, Investigation, Writing—Review & Editing. James Bradley: Conceptualization, Methodology, Investigation, Writing—Original Draft; Writing—Review & Editing. Dmitriy Korobskiy: Methodology, Writing—Review & Editing, Resources. Tandy Warnow: Conceptualization, Methodology, Writing—Original Draft, Writing—Review & Editing. George Chacko: Conceptualization, Methodology, Investigation, Writing—Original Draft, Writing—Review & Editing, Funding Acquisition, Resources, Supervision.

COMPETING INTERESTS

The authors have no competing interests. Scopus data used in this study was available through a collaborative agreement with Elsevier on the ERNIE project. Elsevier personnel played no role in conceptualization, experimental design, review of results, or conclusions presented. The content of this publication is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or Elsevier. Sitaram Devarakonda's present affiliation is Randstad USA. His contributions to this article were made while he was a full-time employee of NET ESolutions Corporation.

SUPPORTING INFORMATION

Supplementary material on K-L calculations is available on our Github site (Korobskiy et al., 2019).

FUNDING INFORMATION

Research and development reported in this publication was partially supported by federal funds from the National Institute on Drug Abuse (NIDA), National Institutes of Health, U.S. Department of Health and Human Services, under Contract Nos. HHSN271201700053C (N43DA-17-1216) and HHSN271201800040C (N44DA-18-1216). Tandy Warnow receives funding from the Grainger Foundation.

DATA AVAILABILITY

Access to the bibliographic data analyzed in this study requires a license from Elsevier. Code generated for this study is freely available from our Github site (Korobskiy et al., 2019).

REFERENCES

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*(1), 47–97. https://doi.org/10.1103/RevModPhys.74.47

Barber, B. (1961). Resistance by scientists to scientific discovery. *Science*, *134*, 596–602. https://doi.org/10.1126/science.134.3479.596

Becke, A. D. (1993). Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics*, *98*(7), 5648–5652. https://doi.org/10.1063/1.464913

Bornmann, L., Leydesdorff, L., & Mutz, R. (2013). The use of percentiles and percentile rank classes in the analysis of bibliometric

data: Opportunities and limits. *Journal of Informetrics*, *7*(1), 158–165. https://doi.org/10.1016/j.joi.2012.10.001

Bornmann, L., Ye, A. Y., & Ye, F. Y. (2018). Identifying "hot papers" and papers with "delayed recognition" in large-scale datasets by using dynamically normalized citation impact scores. *Scientometrics*, *116*(2), 655–674. https://doi.org/10.1007/s11192-018-2772-0

Boyack, K., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, *61*(12), 2389–2404. https://doi.org/10.1002/asi.21419

Boyack, K., & Klavans, R. (2014). Atypical combinations are confounded by disciplinary effects. In *International Conference on Science and Technology Indicators* (pp. 49–58). Leiden, Netherlands: CWTS-Leiden University.

Braam, R. R., Moed, H. F., & van Raan, A. F. J. (1991). Mapping of science by combined co-citation and word analysis. I. Structural aspects. *Journal of the American Society for Information Science*, *42*(4), 233–251. https://doi.org/10.1002/(SICI)1097-4571(199105)42:4<233::AID-ASI1>3.0.CO;2-I

Bradford, M. M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry*, *72*, 248–254. https://doi.org/10.1006/abio.1976.9999

Bradley, J., Devarakonda, S., Davey, A., Korobskiy, D., Liu, S., … Chacko, G. (2020). Co-citations in context: Disciplinary heterogeneity is relevant. *Quantitative Science Studies*, *1*(1), 264–276. https://doi.org/10.1162/qss_a_00007

Carlitz, L. (1961). The Staudt-Clausen Theorem. *Mathematics Magazine*, *34*, 131–146. https://doi.org/10.2307/2688488

Clausen, T. (1840). Theorem. *Astronomische Nachrichten*, *17*, 351–352.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-Law Distributions in Empirical Data. *SIAM Review*, *51*(4), 661–703. https://doi.org/10.1137/070710111

Colavizza, G., Boyack, K., van Eck, N. J., & Waltman, L. (2018). The Closer the Better: Similarity of Publication Pairs at Different Cocitation Levels. *Journal of the Association for Information Science and Technology*, *69*(4), 600–609. https://doi.org/10.1002/asi.23981

Cole, S. (1970). Professional standing and the reception of scientific discoveries. *American Journal of Sociology*, *76*(2), 286–306. Retrieved from https://www.jstor.org/stable/2775594

Crane, D. (1972). *Invisible colleges: Diffusion of knowledge in scientific communities*. Chicago: University of Chicago Press.

Drude, P. (1900a). Zur Elektronentheorie der Metalle. *Annalen der Physik*, *306*, 566–613. https://doi.org/10.1002/andp.19003060312

Drude, P. (1900b). Zur Elektronentheorie der Metalle; II. Teil. Galvanomagnetische und thermomagnetische Effecte. *Annalen der Physik*, *308*, 369–402. https://doi.org/10.1002/andp.19003081102

Elsevier BV. (2019). *Scopus*. Retrieved from https://www.scopus.com/home.uri (accessed December 2019)

Eom, Y.-H., & Fortunato, S. (2011). Characterizing and modeling citation dynamics. *PLOS ONE*, *6*(9), 1–7. https://doi.org/10.1371/journal.pone.0024926

Garfield, E. (1970). Would Mendel's work have been ignored if the Science Citation Index was available 100 years ago? *Essays of an Information Scientist*, *1*, 69–70.

Garfield, E. (1980). Premature Discovery or Delayed Recognition—Why? *Essays of an Information Scientist*, *4*, 488–493.

Glänzel, W., & Garfield, E. (2004). The myth of delayed recognition. *Scientist*, *18*(11).

Gmür, M. (2003). Co-citation analysis and the search for invisible colleges: A methodological evaluation. *Scientometrics*, *57*(1), 27–57. https://doi.org/10.1023/A:1023619503005

Goldstein, M. L., Morris, S. A., & Yen, G. G. (2004). Problems with fitting to the power-law distribution. *The European Physical Journal B—Condensed Matter and Complex Systems*, *41*(2), 255–258. https://doi.org/10.1140/epjb/e2004-00316-5

Gómez, I., Bordons, M., Fernàndez, M., & Méndez, A. (1996). Coping with the problem of subject classification diversity. *Scientometrics*, *35*, 223–235. https://doi.org/10.1007/BF02018480

Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying Sleeping Beauties in science. *Proceedings of the National Academy of Sciences*, *112*(24), 7426–7431. https://doi.org/10.1073/pnas.1424329112

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

Klavans, R., & Boyack, K. (2017). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology*, *68*, 984–998. https://doi.org/10.1002/asi.23734

Korobskiy, D., Davey, A., Liu, S., Devarakonda, S., & Chacko, G. (2019). *Enhanced Research Network Informatics Environment (ERNIE)* (Github Repository). NET ESolutions Corporation. Retrieved from https://github.com/NETESOLUTIONS/ERNIE

Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*, *227*(5259), 680–685. https://doi.org/10.1038/227680a0

Lee, C., Yang, W., & Parr, R. G. (1988). Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B*, *37*(2), 785–789. https://doi.org/10.1103/PhysRevB.37.785

Li, J., & Ye, F. Y. (2016). Distinguishing sleeping beauties in science. *Scientometrics*, *108*(2), 821–828. https://doi.org/10.1007/s11192-016-1977-3

Marshakova-Shaikevich, I. (1973). System of document connections based on references. *Scientific and Technical Information Serial of VINITI*, *6*(2), 3–8. Retrieved from http://garfield.library.upenn.edu/marshakova/marshakovanauchtechn1973.pdf

Merton, R. (1963). Resistance to the systematic study of multiple discoveries in science. *European Journal of Sociology*, *4*(2), 237–282.

Merton, R. (1968). The Matthew effect in science. *Science*, *159*(3810), 56–63.

Milojevic, S. (2019). Practical method to reclassify web of science articles into unique subject categories and broad disciplines. *Quantitative Science Studies*, *1*(1), 183–206. https://doi.org/10.1162/qss_a_00014

Mitzenmacher, M. (2003). A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, *1*, 226–251.

Montebruno, P., Bennett, R. J., van Lieshout, C., & Smith, H. (2019). A tale of two tails: Do power law and lognormal models fit firm-size distributions in the mid-Victorian era? *Physica A: Statistical Mechanics and its Applications*, *523*, 858–875. https://doi.org/10.1016/j.physa.2019.02.054

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, *45*(2), 167–256. https://doi.org/10.1137/S003614450342480

Noma, E. (1984). Co-citation analysis and the invisible college. *Journal of the American Society for Information Science*, *35*(1), 29–33. https://doi.org/10.1002/asi.4630350105

Perline, R. (2005). Strong, weak and false inverse power laws. *Statistical Science*, *20*(1), 68–88.
Press, W., Teukolsky, S., Vetterling, W., & Flannery, B. (2007). *Numerical recipes in C: The art of scientific computing* (3rd ed.). New York: Cambridge University Press.
Price, D. de Solla. (1965). Networks of Scientific Papers. *Science*, *149*, 510–515.
Price, D. de Solla. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*(5), 292–306. https://doi.org/10.1002/asi.4630270505
Price, D. de Solla, & Beaver, D. D. (1966). Collaboration in an invisible college. *American Psychologist*, *21*(11), 1011–1018.
Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, *105*(45), 17268–17272.
Redner, S. (2005). Citation statistics from 110 years of physical review. *Physics Today*, *58*(6), 49–54.
Shu, F., Julien, C.-A., Zhang, L., Qiu, J., Zhang, J., & Larivière, V. (2019). Comparing journal and paper level classifications of science. *Journal of Informetrics*, *13*(1), 202–225. https://doi.org/10.1016/j.joi.2018.12.005
Small, H. (1973). Cocitation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*(4), 265–269. https://doi.org/10.1002/asi.4630240406
Small, H., & Greenlee, E. (1980). Citation context analysis of a co-citation cluster: Recombinant-DNA. *Scientometrics*, *2*(4), 277–301. https://doi.org/10.1007/BF02016349
Small, H., & Sweeney, E. (1985). Clustering the science citation index® using cocitations. *Scientometrics*, *7*(3), 391–409. https://doi.org/10.1007/BF02017157
Small, H., Sweeney, E., & Greenlee, E. (1985). Clustering the science citation index using co-citations. II. Mapping science. *Scientometrics*, *8*(5), 321–340. https://doi.org/10.1007/BF02018057
StackExchange. (2014). *Can I use the Kolmogorov–Smirnov test on my data?* Retrieved from https://stats.stackexchange.com/questions/112910/can-i-use-kolmogorov-smirnov-test-on-my-data
Stringer, M. J., Sales-Pardo, M., & Amaral, L. A. N. (2008). Effectiveness of journal ranking schemes as a tool for locating information. *Journal of the American Society for Information Science and Technology*, *3*(2), e1683.
Stringer, M. J., Sales-Pardo, M., & Amaral, L. A. N. (2010). Statistical validation of a global model for the distribution of the ultimate number of citations accrued by papers published in a scientific journal. *PLOS ONE*, *61*(7), 1377–1385.
Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical Combinations and Scientific Impact. *Science*, *342*(6157), 468–472. https://doi.org/10.1126/science.1240474
van Raan, A. F. J. (1990). Fractal dimension of co-citations. *Nature*, *347*(6294), 626. https://doi.org/10.1038/347626a0
van Raan, A. F. J. (2004). Sleeping Beauties in science. *Scientometrics*, *59*(3), 467–472. https://doi.org/10.1023/B:SCIE.0000018543.82441.f1
van Raan, A. F. J., & Winnink, J. J. (2019). The occurrence of 'sleeping beauty' publications in medical research: Their scientific impact and technological relevance. *PLOS ONE*, *14*(10), 1–34. https://doi.org/10.1371/journal.pone.0223373
von Staudt, K. (1840). Beweis eines Lehrsatzes, die Bernoullischen Zahlen betreffend. *Journal für die reine und angewandte Mathematik*, *21*, 372–374.
Wallace, M., Larivière, V., & Gingras, Y. (2009). Modeling a century of citation distributions. *Journal of Informetrics*, *3*(4), 296–303.
Waltman, L., & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378–2392. https://doi.org/10.1002/asi.22748
Wang, D., Song, C., & Barabási, A.-L. (2013). Quantifying Long-Term Scientific Impact. *Science*, *342*(6154), 127–132. https://doi.org/10.1126/science.1237825
Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, *46*(8), 1416–1436. https://doi.org/10.1016/j.respol.2017.06.006