Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Citation counts, distinct citing authors, and concentration:a mathematical study of the ratio D= C/U

Alberto Verjovsky

albertoverjovsky@gmail.com

Universidad Nacional Autónoma de México

Research Article

Keywords:

Posted Date: November 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11556999/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Citation counts, distinct citing authors, and concentration: a mathematical study of the ratio $D = C/U$

Alberto Verjovsky
Instituto de Matemáticas, Unidad Cuernavaca
Universidad Nacional Autónoma de México
albertoverjovsky@gmail.com

October 29, 2026

Abstract

Let C be the number of retained citation links to a fixed collection of publications and U the number of distinct authors of the citing papers. The ratio $D = C/U$ is the mean citation credit when each link is divided equally among its citing authors. Its relation to repetition depends on team size: $D = R/\overline{k}$, where R is the mean number of author–citation incidences and $\overline{k}$ is the mean team size weighted by citation links. Finite counting arguments give sharp bounds for concentration indices, an expression for concentration in terms of overlap between citing teams, and formulas for the changes caused by additional citations and by errors in author identification. Subsampling identities permit comparisons at equal numbers of links or papers. A public-data pilot examines six articles from one volume of *Geometry & Topology*. After version consolidation and screening, 89 citing works remain, of which 75 have no identified author in common with the target paper. Two incorrect citation links and nine works with incomplete author identifiers show the limitations of the retrieved records. The results distinguish mean credit, concentration, and the number of distinct citing authors; the pilot is too small to say how these quantities behave across fields.

Keywords: bibliometrics; citing authors; citation concentration; fractional counting; author identification.

1 Introduction and preceding work

A citation count does not specify how many people appear as authors of the citing publications. One hundred citation links can come from one hundred single-author papers written by different people, or from one hundred papers written by the same person, with each paper contributing one link. If C denotes the number of citation links and U the number of distinct citing authors, the ratio

$$D = \frac{C}{U} \tag{1}$$

distinguishes these two examples: its values are 1 and 100, respectively. This observation motivates the study of D together with its numerator and denominator.

The use of citing authors in bibliometrics has substantial precedents. Ajiferuke and Wolfram (2010) studied citer analysis and a citer-based analogue of the h-index, the ch-index, which Franceschini et al. (2010) analysed further. A portfolio of papers has ch-index n if n is the largest integer such that at least n of its papers each have at least n distinct citing authors; citing authors are counted separately for each cited paper.

Jovanović and Fritsche (2011) introduced the NUCA family of indicators. Their construction distinguishes repeated citing-author occurrences from distinct citing authors associated with each cited publication, so the same person can contribute again by citing a different publication in the evaluated portfolio. This differs from a union of citing authors over the entire portfolio. They also discuss the effect of large author lists.

Cappelletti-Montano et al. (2021) introduced a citing-authors index and studied it using MathSciNet data from pure mathematics. Their index is $\max\{0, N_a - N_{\mathrm{co}}\}$, where N_a counts distinct authors citing at least one publication of the evaluated researcher, and N_{co} counts that researcher's distinct coauthors, including the researcher. This portfolio-level count is a close predecessor of U. Subtracting the size of a coauthor set is not the same operation as removing links with shared authors, so values can be compared only after the exclusion rules are matched.

At the journal level, Rowlands (2002) introduced a diffusion factor, originally expressed per 100 citations. In the formulation studied by Egghe (2005), without the factor 100, it is the number of distinct citing journals divided by the number of citations, for fixed citing and cited periods. Replacing journals by people and inverting this ratio leads formally to (1); the difference is that a citing paper belongs to one journal but can have many authors.

These measures use different counting units: authors associated with each target paper, authors counted once over a portfolio, and journals containing citing papers. The ratio D uses a portfolio-wide union of citing authors and keeps the number of citation links as its numerator. To the best of our knowledge, the following results are new in this setting: the decomposition $D = R/\overline{k}$ (Proposition 3.1), the expression of concentration through overlap of citing teams (Proposition 5.2), the effect of overlapping citing-author sets on aggregation (Proposition 6.2), and the bounds for identification errors and missing papers (Section 8). The proofs use finite counting together with standard inequalities for concentration indices. The empirical example shows how version consolidation and erroneous reference links change these quantities.

A recent conference abstract by Evdaimon et al. (2025) examines citation concentration through, among other quantities, the number of citing authors accounting for half of an author's citations. This is a further precedent for studying the share of citations coming from small sets of citing authors. Our subset bounds (Proposition 5.1) concern a normalized fractional distribution; comparing them with empirical counts of this kind requires treating jointly written papers in the same way.

2 Citation links and author counts

Fix a target author A, a finite set P of publications authored by A, a period for citing publications, and a database snapshot. A *citation link* is an ordered pair (q, p), where $p \in P$ and the publication q cites p. A reference repeated several times within q contributes one link to p. A paper citing three different members of P contributes three links. Versions identified as the same work are consolidated before counting, and the author set of a consolidated work is taken from one designated version.

We consider two self-citation rules. The *author-exclusion rule*, used in the mathematical sections unless stated otherwise, discards every citing publication with A among its authors; a paper is not discarded merely because one of its authors has collaborated with A. The stricter *shared-author rule* removes each citation link (q, p) for which q and p have an author in common; all counts are then recomputed from the retained links. Since A is an author of every $p \in P$, under either rule A is not among the citing authors.

Let L be the finite set of links remaining after consolidation and the chosen exclusions,

and let

$$Q = \{q : \text{there exists } p \in P \text{ with } (q,p) \in L\}.$$

For each $q \in Q$, let S_q be its finite, nonempty set of distinct authors. Define

$$r_q = |\{p \in P : (q,p) \in L\}|, \qquad k_q = |S_q|,$$

where $|X|$ denotes the cardinality of a finite set X. Thus $1 \le r_q \le |P|$ and $k_q \ge 1$. Author identities are assumed resolved in the mathematical model. Records with absent or incomplete author lists require separate reporting, as discussed in Section 8.

Define the set of citing authors and the principal counts by

$$V = \bigcup_{q \in Q} S_q, \quad U = |V|, \quad N = |Q|, \quad C = \sum_{q \in Q} r_q, \quad T = \sum_{q \in Q} r_q k_q. \tag{2}$$

Here N counts citing papers, C counts citation links, and T counts author–citation incidences. An incidence is a triple (i,q,p) with $i \in S_q$ and (q,p) a retained link.

For $i \in V$, put

$$t_i = \sum_{q : i \in S_q} r_q.$$

Since $S_q \subseteq V$ for every $q \in Q$, interchanging finite sums gives

$$\sum_{i \in V} t_i = \sum_{q \in Q} r_q \sum_{i \in V} \mathbf{1}_{\{i \in S_q\}} = \sum_{q \in Q} r_q k_q = T.$$

Each $i \in V$ belongs to a paper with a retained link, so $1 \le t_i \le C$. The integer t_i counts the citation links attached to papers bearing i's name, not the references that i personally chose; throughout, "citing author" refers to authorship records in this sense.

When $C = 0$, the sets L, Q, V are empty and D is left undefined. Unless explicitly stated otherwise, all subsequent ratios and concentration quantities concern a nonempty retained link set, so C, U, T are positive. Sums indexed by authors are over V.

Example 2.1 (A multi-author citation)**.** One paper with ten authors cites one target paper. Then $C = 1$, $U = 10$, $T = 10$, and $D = 1/10$. Each author has exactly one citation incidence. Thus D is not in general the average number of citation incidences per author, and it is not bounded below by 1.

Example 2.2 (Several references in one paper)**.** Suppose $|P| \ge 20$. One single-author paper cites twenty distinct members of P. Then $N = U = 1$ and $C = T = 20$. The ratio $D = 20$ describes twenty links in one paper, not twenty separate citing papers; the number of citing papers per citing author is $N/U = 1$.

3 Repetition and citing-team size

Define

$$R = \frac{T}{U}, \qquad \overline{k} = \frac{T}{C}.$$

The number R is the mean number of citation incidences per citing author. The number $\overline{k}$ is the mean citing-team size when each citation link has equal weight: a paper with r_q target references appears r_q times in this mean.

Proposition 3.1. *For every nonempty retained citation collection,*

$$D = \frac{R}{\overline{k}}. \tag{3}$$

Moreover $R \geq 1$, with equality exactly when every citing author has one incidence. Always $D \leq C$. If $K \geq 1$ is an integer and $k_q \leq K$ for all q, then

$$\frac{1}{K} \leq D \leq C. \tag{4}$$

For every pair of positive integers C, K, both bounds are attained by some citation collection.

Proof. Since $T > 0$, one has $(T/U)/(T/C) = C/U$. Every t_i is a positive integer, so $T \geq U$, with equality exactly when all $t_i = 1$. Since $U \geq 1$, $D \leq C$. Also $U \leq T \leq KC$, which gives $D \geq 1/K$. For the lower bound, take C papers, each citing one target paper, with pairwise disjoint author sets of size K. For the upper bound, take C single-author citing papers written by the same person. □

When all citing papers have the same team size k, one has $\overline{k} = k$ and $R = kD$. Across collections with different team sizes, changes in D can therefore reflect either repetition or coauthorship. In particular, $D = 1$ is not a general criterion for an audience in which every person appears once; the criterion is $R = 1$.

Proposition 3.2 (Information absent from paper citation counts)**.** *Fix integers $c_p \geq 0$ for $p \in P$, with $C = \sum_p c_p > 0$. There exist single-author citing collections in which p receives exactly c_p links for each p, with $U = 1$, and others with $U = C$. These collections have the same total citations and h-index, but their values of D are C and 1.*

Proof. Create c_p distinct citing papers for each p, each citing only p among the members of P. Then p receives c_p links. Assign all these papers to one author different from A, or assign each to a different such author. The vector (c_p) is the same in both cases. The h-index, the largest integer h such that at least h target papers have at least h citations each (Hirsch, 2005), depends only on that vector. □

4 Fractional allocation and concentration

Allocate one unit of credit for each citation link equally among the authors of its citing paper; this is an accounting convention, not a claim that coauthors contributed equally to the choice of reference. Define

$$x_i = \sum_{q : i \in S_q} \frac{r_q}{k_q}, \qquad s_i = \frac{x_i}{C} \quad (i \in V). \tag{5}$$

Each x_i is positive, and finite summation gives

$$\sum_{i \in V} x_i = \sum_{q \in Q} \frac{r_q}{k_q} |S_q| = C, \qquad \sum_{i \in V} s_i = 1.$$

Thus s_i is the share of citation credit allocated to author i, and D is the mean fractional credit:

$$D = \frac{1}{U} \sum_{i \in V} x_i. \tag{6}$$

This holds for arbitrary team sizes. The mean does not determine how the credit is distributed among authors.

Define

$$H = \sum_{i \in V} s_i^2, \qquad E_2 = \frac{1}{H}. \tag{7}$$

The quantity H is the Simpson concentration, also known as the Herfindahl concentration. Its reciprocal E_2, the inverse Simpson effective number, is the number of equally weighted authors that would give the same concentration; it need not be an integer. It is the Hill number of order two (Hill, 1973).

Theorem 4.1. *Let $\sigma^2 = U^{-1} \sum_{i \in V} (x_i - D)^2$. Then*

$$H = \frac{1}{U}\left(1 + \frac{\sigma^2}{D^2}\right), \qquad E_2 = \frac{U}{1 + \sigma^2/D^2}. \tag{8}$$

In particular, $1 \le E_2 \le U$, and $E_2 = U$ if and only if all x_i are equal. If $U > 1$, then $E_2 > 1$.

Proof. Since $\sum_i x_i = UD$, expansion gives

$$U\sigma^2 = \sum_i x_i^2 - 2D \sum_i x_i + UD^2 = \sum_i x_i^2 - UD^2.$$

Divide by $C^2 = U^2 D^2$ to obtain (8). Since $\sigma^2 \ge 0$, $E_2 \le U$, with equality exactly when each square $(x_i - D)^2$ vanishes. Since the s_i are positive and sum to 1,

$$H = \sum_i s_i^2 \le \Big(\sum_i s_i\Big)^2 = 1,$$

so $E_2 \ge 1$. If $U > 1$, the expansion of $(\sum_i s_i)^2$ contains at least one positive cross term $2 s_i s_j$, so $H < 1$ and $E_2 > 1$. □

Example 4.2 (Identical D, different concentration)**.** Consider 100 single-author papers, each making one target citation, and ten distinct citing authors. Two such incidence vectors are

$$(10, 10, 10, 10, 10, 10, 10, 10, 10, 10), \qquad (91, 1, 1, 1, 1, 1, 1, 1, 1, 1).$$

Both give $C = 100$, $U = 10$, and $D = 10$. In the first case $H = 0.1$ and $E_2 = 10$. In the second,

$$H = \frac{8290}{10000} = 0.829, \qquad E_2 = \frac{10000}{8290} \approx 1.2063.$$

The ratio D does not distinguish uniform repetition from a distribution in which one author receives most of the credit.

Proposition 4.3 (Sharp single-author bounds)**.** *Suppose every citing publication has one author. Write $C = aU + b$, where $a = \lfloor C/U \rfloor$ and $0 \le b < U$. Then*

$$\frac{(U-b)a^2 + b(a+1)^2}{C^2} \le H \le \frac{(C - U + 1)^2 + U - 1}{C^2}. \tag{9}$$

The lower bound is attained when $U - b$ authors have a links and b authors have $a + 1$ links. The upper bound is attained when one author has $C - U + 1$ links and the others have one each.

Proof. Here $k_q = 1$ for all q, so $x_i = t_i$, and these are positive integers summing to C; in particular $C \ge U \ge 1$ and $a \ge 1$. Since $H = C^{-2} \sum_i x_i^2$, it suffices to bound the sum of squares over all vectors of U positive integers summing to C. This set is finite and nonempty, so the minimum and maximum are attained.

If $x_i \geq x_j + 2$, moving one unit from x_i to x_j keeps all entries positive and decreases the sum of squares by $2(x_i - x_j - 1) > 0$. Hence at a minimizer all entries lie in $\{c, c+1\}$ for some integer c, with at least one entry equal to c. Then $cU \leq C < (c+1)U$, so $c = a$, and the number of entries equal to $a+1$ is $C - aU = b$. This gives the lower bound.

If $x_i \geq x_j \geq 2$, moving one unit from x_j to x_i keeps all entries positive and increases the sum of squares by $2(x_i - x_j + 1) > 0$. Hence a maximizer has at most one entry greater than 1; since the entries sum to C, that entry is $C - U + 1$. This gives the upper bound.

Each extremal vector is realized by distinct single-author citing papers, each with one target link, so both bounds are attained among collections with the given C and U. If the citing papers are fixed in advance, the bounds need not be attained. □

Remark 4.4. The single-author hypothesis is essential, because it makes each x_i an integer. In Example 2.1, all $x_i = 1/10$ and $C < U$, so the integer exchange argument does not apply to multi-author credits.

The order-one effective number is

$$E_1 = \exp\left(-\sum_{i \in V} s_i \log s_i\right),$$

where log is the natural logarithm; it is defined because every s_i is positive. One has $E_2 \leq E_1 \leq U$. Indeed, Jensen's inequality for the concave function log, with weights s_i, gives

$$\sum_i s_i \log s_i \leq \log\left(\sum_i s_i^2\right), \qquad \sum_i s_i \log(1/s_i) \leq \log\left(\sum_i s_i / s_i\right) = \log U,$$

and exponentiating gives the two inequalities. These are standard diversity inequalities, applied here to the distribution (5).

5 Subsets of citing authors and shared papers

For $B \subseteq V$, let

$$m = |B|, \qquad \alpha_B = \sum_{i \in B} s_i.$$

Thus α_B is the fraction of citation credit allocated to B.

Proposition 5.1. *If $1 \leq m < U$, then*

$$H \geq \frac{\alpha_B^2}{m} + \frac{(1 - \alpha_B)^2}{U - m}, \tag{10}$$

with equality exactly when the shares are constant on B and constant on $V \setminus B$. For every $B \subseteq V$,

$$m \geq \alpha_B^2 E_2, \tag{11}$$

and the inequality is strict when B is nonempty and $B \neq V$.

Proof. Let $1 \leq m < U$. By the Cauchy–Schwarz inequality applied separately to B and to $V \setminus B$,

$$\alpha_B^2 \leq m \sum_{i \in B} s_i^2, \qquad (1 - \alpha_B)^2 \leq (U - m) \sum_{i \notin B} s_i^2,$$

with equality exactly when the shares are constant on the respective set. Dividing by m and by $U - m$ and adding gives (10) and its equality case. Since $1 - \alpha_B > 0$, (10) gives $H > \alpha_B^2/m$, which is the strict form of (11). If $B = \varnothing$, both sides of (11) are 0. If $B = V$, (11) reads $U \geq E_2$, which is Theorem 4.1. □

For example, if $E_2 = 100$, a proper subset receiving at least half the fractional credit has more than 25 authors, hence at least 26, while $U \geq 100$.

There is a useful paper-level formula. Define

$$w_q = \frac{r_q}{C}, \qquad \kappa(q, q') = \frac{|S_q \cap S_{q'}|}{k_q k_{q'}} \quad (q, q' \in Q).$$

The number $\kappa(q, q')$ is the probability that independent uniform choices of one author from each of the two papers agree. In particular $\kappa(q, q) = 1/k_q$.

Proposition 5.2. *The concentration is*

$$H = \sum_{q,q' \in Q} w_q w_{q'} \kappa(q, q'). \tag{12}$$

Consequently,

$$H \geq \sum_{q \in Q} \frac{w_q^2}{k_q}, \tag{13}$$

with equality exactly when distinct citing papers have disjoint author sets.

Proof. Equation (5) gives

$$s_i = \sum_{q \in Q} w_q \frac{\mathbf{1}_{\{i \in S_q\}}}{k_q}.$$

Square, sum over $i \in V$, and interchange finite sums, using $\sum_{i \in V} \mathbf{1}_{\{i \in S_q\}} \mathbf{1}_{\{i \in S_{q'}\}} = |S_q \cap S_{q'}|$. This gives (12). Its diagonal terms are w_q^2/k_q. Every off-diagonal term is nonnegative, and, since $w_q, w_{q'} > 0$, it vanishes exactly when $S_q \cap S_{q'} = \varnothing$. Equality in (13) is therefore equivalent to pairwise disjoint author sets. □

If every citing paper is written by the same team of k people, then $s_i = 1/k$ and $E_2 = k$, regardless of the number of papers. The same fractional distribution can arise from different authorship patterns; Table 1 includes such a comparison.

6 Additional citations and aggregation

Proposition 6.1 (Adding citing papers)**.** *Keep the existing retained links, author identities, and exclusion rule fixed. Add distinct citing works contributing $\Delta C > 0$ new retained links. Let $v \geq 0$ be the number of distinct authors in the added works who were not previously in V. If D' is the resulting ratio, then*

$$D' - D = \frac{U\Delta C - Cv}{U(U + v)}. \tag{14}$$

For $v > 0$, the ratio decreases exactly when $\Delta C/v < D$, and is unchanged exactly when $\Delta C/v = D$. For $v = 0$ it increases.

Proof. The new counts are $C + \Delta C$ and $U + v$, and $(C + \Delta C)/(U + v) - C/U$ equals the right side of (14). □

For one new citing paper with r retained target links and v new authors, the condition for a decrease is $r < vD$. Thus additional citations can decrease D.

For aggregation, partition the citation links into two nonempty disjoint collections, with counts C_1, C_2, citing-author sets V_1, V_2, $U_j = |V_j|$, and ratios $D_j = C_j/U_j$ for $j = 1, 2$. Such a partition can arise from disjoint time periods or disjoint subsets of target papers. Let $J = |V_1 \cap V_2|$. A citing paper with links in both collections contributes its whole team to $V_1 \cap V_2$.

Proposition 6.2 (Overlap of citing-author sets)**.** *For the union of the two link collections,*

$$D = \frac{C_1 + C_2}{U_1 + U_2 - J} = \frac{U_1 D_1 + U_2 D_2}{U_1 + U_2 - J}. \tag{15}$$

In particular, D is at least the U-weighted mean of D_1 and D_2:

$$D \geq \frac{U_1 D_1 + U_2 D_2}{U_1 + U_2},$$

with equality exactly when $J = 0$.

Proof. The citation counts add because the link sets are disjoint, and inclusion–exclusion gives $|V_1 \cup V_2| = U_1 + U_2 - J$. The inequality and its equality case follow because $C_1 + C_2 > 0$. □

When portfolios of different researchers are aggregated, two further issues arise. A citation to a jointly authored target paper may be counted twice unless the link sets are made disjoint. Also, the author-exclusion rule is defined relative to one target author, so the retained links of the combined portfolio need not be the union of the separately retained links. At the level of an institution, one must first decide whether the objects counted are distinct works, author–work pairs, or another explicitly stated unit.

Proposition 6.3 (Replication and scale)**.** *Let $\ell \geq 1$ be an integer. Form ℓ copies of a citation configuration, using distinct eligible citing works that remain distinct after version consolidation, but the same citing-author identities. Then D and R are multiplied by ℓ, while H and E_2 are unchanged. If the ℓ copies instead use mutually disjoint copies of the citing-author set, then D and R are unchanged, while H is divided by ℓ and E_2 is multiplied by ℓ.*

Proof. In the first construction, C, T and each x_i are multiplied by ℓ and U is unchanged, so the shares s_i are unchanged. In the second, C, T, U are all multiplied by ℓ, and the shares consist of ℓ copies of $(s_i/\ell)_{i \in V}$. The assertions follow directly. □

Thus D does not measure audience size: one citation from one person and one thousand citations from one thousand people both give $D = 1$.

7 Sampling and comparison at equal citation volume

A cumulative ratio depends on the amount of material observed, even when the citing distribution is fixed. Let $M \geq 1$ be an integer. Consider independent, identically distributed single-author citation events, where author i is chosen with probability π_i, for $1 \leq i \leq M$, with $\pi_i > 0$ and $\sum_i \pi_i = 1$. Let U_n be the number of distinct authors among an integer number $n \geq 1$ of events. Each event is a single-author paper with one target link, so n/U_n is the value of D for the first n events; the model does not describe multi-author papers.

Proposition 7.1 (Expected number of citing authors)**.** *In this model,*

$$\mathbb{E}U_n = \sum_{i=1}^{M} [1 - (1 - \pi_i)^n]. \tag{16}$$

For uniform probabilities, this becomes $M[1 - (1 - 1/M)^n]$. Moreover,

$$\mathbb{E}\left[\frac{n}{U_n}\right] \geq \frac{n}{\mathbb{E}U_n}. \tag{17}$$

Proof. Write U_n as the sum of the indicators that each author appears. The probability that i never appears is $(1-\pi_i)^n$. Taking expectations proves (16). Since $1 \le U_n \le \min(n, M)$, all expectations are finite, and Jensen's inequality for the convex function $u \mapsto 1/u$ on $(0, \infty)$ proves (17). □

Since $U_n \le M$, one has $n/U_n \ge n/M$, so the ratio tends to infinity even though the citing probabilities remain fixed. It need not increase at each step: the sequence of authors $1, 1, 2$ gives ratios $1, 2, 3/2$. The quotient $n/\mathbb{E}U_n$ is not in general the expectation of n/U_n.

For a finite observed collection, an exact comparison is possible by subsampling. Let $1 \le n \le C$, and choose uniformly a subset of n distinct citation links. For author i, the number of links whose citing papers include i is t_i. The expected number of distinct citing authors in the selected links is

$$U^{\text{link}}(n) = \sum_{i \in V} \left[1 - \frac{\binom{C-t_i}{n}}{\binom{C}{n}} \right], \tag{18}$$

where $\binom{c}{n} = 0$ when $n > c$. To prove the formula, note that author i is absent exactly when the subset avoids all t_i links of i, count such subsets, and sum the presence probabilities. No independence between coauthors is needed.

Sampling whole citing papers gives a different comparison. Let $\nu_i = |\{q \in Q : i \in S_q\}|$ be the number of citing papers bearing i's name. For $1 \le n \le N$, the expected number of distinct authors in n uniformly sampled papers is

$$U^{\text{paper}}(n) = \sum_{i \in V} \left[1 - \frac{\binom{N-\nu_i}{n}}{\binom{N}{n}} \right]. \tag{19}$$

These are finite-population counting identities. They standardize the number of observed links or papers; without additional assumptions they do not correct for field differences or estimate unobserved authors.

In the independent single-author model, if n_i is the observed count of author i and $n \ge 2$, then

$$\widehat{H} = \frac{\sum_i n_i(n_i - 1)}{n(n-1)}$$

is an unbiased estimator of $\sum_i \pi_i^2$: there are $n(n-1)$ ordered pairs of different events, and each pair agrees in author with probability $\sum_i \pi_i^2$. The argument does not apply to lists of coauthors treated as if they were independent citation events. The reciprocal $1/\widehat{H}$ is undefined on samples with no repeated author, which have positive probability when $n \le M$; in that case $\mathbb{E}[1/\widehat{H}]$ is not defined. Even when it is defined, it is biased in general. For two equally likely authors and $n = 3$, it equals 1 with probability $1/4$ and 3 with probability $3/4$, so its expectation is $5/2$ rather than 2. For $M = 1$, it equals 1 almost surely.

8 Author identification and incomplete records

Author disambiguation assigns records to people. Splitting one person across several identities increases the recorded number of citing authors; merging different people decreases it. Even when the citation count is correct, either error changes D.

Proposition 8.1 (Identity-count errors)**.** *Suppose C is unchanged and the recorded number of citing authors is $U + \delta > 0$. Then*

$$\frac{D_{\text{recorded}}}{D} = \frac{U}{U+\delta}, \qquad \frac{D_{\text{recorded}} - D}{D} = -\frac{\delta}{U+\delta}.$$

If $0 \le \varepsilon < U$ and $|\delta| \le \varepsilon$, the absolute relative error is at most $\varepsilon/(U-\varepsilon)$.

Proof. Substitution gives the identities. Since $U + \delta \geq U - \varepsilon > 0$,

$$\left| \frac{D_{\text{recorded}} - D}{D} \right| = \frac{|\delta|}{U + \delta} \leq \frac{\varepsilon}{U - \varepsilon}.$$

□

The hypothesis that C is unchanged can fail. If an identification error involves the target author's own identity, it changes which papers are excluded as self-citations, and hence C; the pilot in Section 10 contains such a case.

At the level of the distribution, merging two identities with shares u and v into one identity with share $u + v$ increases H by

$$(u + v)^2 - u^2 - v^2 = 2uv,$$

and the corresponding split decreases it by the same amount. This calculation assumes that the merged identities never appear on the same paper. If they do, the merge changes the team size k_q of that paper, the allocation (5) must be recomputed, and the formula $2uv$ no longer applies.

One can also bound the effect of missing citing papers without imputing identities. Suppose the observed collection has $C_o > 0$ links and $U_o > 0$ authors. Keep the observed works, retained links, author identities, and exclusion rule fixed. Suppose that $\mu \geq 0$ additional distinct eligible citing works are missing, each with between 1 and Λ retained target links and at most K' authors, where Λ, K' are positive integers. Then the completed ratio satisfies

$$\frac{C_o + \mu}{U_o + K'\mu} \leq D_{\text{complete}} \leq \frac{C_o + \Lambda\mu}{U_o}. \tag{20}$$

Indeed, the numerator lies between $C_o + \mu$ and $C_o + \Lambda\mu$, while the denominator lies between U_o and $U_o + K'\mu$. Since a citing work has at most $|P|$ target links, Λ can be replaced by $\min(\Lambda, |P|)$. With this replacement both bounds are attained: the lower one when every missing work has one link and K' authors not otherwise present, the upper one when every missing work has $\min(\Lambda, |P|)$ links and only authors already in the observed collection. Without bounds on the number of missing papers and their team sizes, the observed ratio gives no such guarantee.

OpenAlex provides authorship and reference information for each work (OpenAlex, 2026a,b). Its documentation on author disambiguation explicitly describes splitting and merging errors (OpenAlex, 2026c). A reproducible calculation must keep the extraction date, the work and author identifiers returned in that snapshot, and any manual corrections. Incomplete or truncated author lists must not be treated as complete teams.

9 Synthetic examples

Table 1 collects synthetic examples with exact counts; nonterminating values of E_2 are rounded. In every row each citing paper has one target link, so $N = C$. The target portfolio can be taken to consist of one paper; then every row has $h = 1$.

The second and fourth rows have the same C, U, D, E_2 but different repetition R. The first and last rows have the same U and E_2 but very different numbers of citing papers. These examples show that D, R, U, and E_2 carry different information.

Table 1: Synthetic citation collections. No row represents an observed researcher.

Citing authorship pattern	C	U	D	R	E_2
100 different single authors	100	100	1	1	100
10 single authors, 10 papers each	100	10	10	10	10
10 single authors, counts $(91, 1, \ldots, 1)$	100	10	10	10	1.2063
One team of 10, on 100 papers	100	10	10	100	10
100 disjoint teams of 10	100	1000	0.1	1	1000
One paper with 100 authors	1	100	0.01	1	100

10 A public-data pilot

10.1 Selection and data

The sample consists of the first six research articles in *Geometry & Topology*, volume 19 (2015), ordered by first page. The selection rule was recorded before the citing records were inspected and was checked against the publisher's table of contents. Each target is a single paper, so $r_q = 1$ and $C = N$ for each retained collection. These are measurements for individual papers, not for the complete publication records of their authors; in particular no target author A is fixed, and self-citations are handled by the shared-author rule of Section 2 (see below).

P1 L. P. Roberts, *Totally twisted Khovanov homology*, pp. 1–59.

P2 H. N. Sá Earp, *G_2–instantons over asymptotically cylindrical manifolds*, pp. 61–111.

P3 D. Dugger, *Bigraded cohomology of $\mathbb{Z}/2$–equivariant Grassmannians*, pp. 113–170.

P4 M. Golla, *Ozsváth–Szabó invariants of contact surgeries*, pp. 171–235.

P5 M. Deraux and E. Falbel, *Complex hyperbolic geometry of the figure-eight knot*, pp. 237–293.

P6 S. Schreieder, *On the construction problem for Hodge numbers*, pp. 295–342.

The DOI of each target has the form `10.2140/gt.2015.19.`f, where f is its first page. OpenAlex supplied the citing-work records, reference links, and author identifiers in a snapshot retrieved on October 29, 2026. The queries kept citing records dated on or before that day, including preprints and citations made before the journal publication of a target. The saved responses contain 116 target–record pairs. For P3 the query returned 11 records, although the citation count stored in the target's OpenAlex record (the field `cited_by_count`) was 12. The discrepancy was not resolved; the calculations use the retrieved records.

10.2 Consolidation and exclusions

Records were first grouped when they shared a DOI, an arXiv identifier, or an exactly normalized title and author-name list, and these matches were closed transitively. Four further pairs were consolidated after examination of repository or publication information: Charbonneau–Harland's *Deformations of nearly Kähler instantons*; Joyce–Upmeier's *Canonical orientations for moduli spaces of G_2-instantons*; Conway's *Contact surgeries on Legendrian figure-eight knots*; and Parker–Will's *A complex hyperbolic Riley slice*. The evidence and both work identifiers for every manual match are recorded in the supplementary data. In all, 14 redundant records were removed, leaving 102 groups, each representing one citing work.

For each group the metadata of the published article were used when available. When necessary, a complete author-identifier list from another version in the group supplied the authorships, and its source was recorded. A list was accepted only if it was nonempty, had distinct identifiers, contained neither OpenAlex's null identifier nor its placeholder for deleted authors, and had fewer than 100 entries. The last condition excludes lists that may have been truncated, so no team of 100 or more authors was retained. These checks do not verify the identities themselves or that the author list is complete.

Each excluded group was assigned to one category, in this order of precedence: incorrect target link (two groups, for P1 and P6), items without authors such as front matter or editorial notes (two groups, both for P5), and unresolved author identifiers (nine groups). This leaves 89 retained groups. All quantities for a target are computed on the same retained groups. The external subset applies the shared-author rule of Section 2: it removes every group sharing an identified author with its target paper, leaving 75 groups. The all-citations calculation keeps the 14 identified self-citations. Daniel Dugger's missing author identifier on the target record was provisionally resolved using his name, University of Oregon affiliation, and publication list; the original record and the correction are preserved. This is an instance of the identification problem noted after Proposition 8.1: without the correction, the shared-author rule could not have been applied to P3.

Table 2: Screening of target–record pairs. "Other" comprises incorrect links (P1, P6) and items without authors (P5).

Target	Retrieved	Groups	Missing IDs	Other	Retained	External
P1	8	8	1	1	6	5
P2	34	28	4	0	24	19
P3	11	10	0	0	10	10
P4	11	9	1	0	8	8
P5	35	33	1	2	30	22
P6	17	14	2	1	11	11
Total	116	102	9	4	89	75

10.3 Results and sensitivity

Table 3 reports the external subset. Since $C = N$, the two subsampling expectations (18) and (19) coincide; the last column gives their common value $U^{\mathrm{link}}(5)$ at five citing works. Every target has at least five retained external works. P3 and P5 have similar values of D

Table 3: External citing works after screening. Decimals are rounded.

Target	C	U	D	R	E_2	$U^{\mathrm{link}}(5)$
P1	5	8	0.625	1.250	7.14	8.00
P2	19	33	0.576	1.212	20.77	10.07
P3	10	9	1.111	1.778	6.00	5.95
P4	8	9	0.889	1.667	6.16	6.88
P5	22	20	1.100	1.750	13.76	7.00
P6	11	13	0.846	1.385	6.79	6.95

(1.111 and 1.100) and very different values of E_2 (6.00 and 13.76). Most of this difference reflects the number of citing authors ($U = 9$ and $U = 20$): the ratios E_2/U are 0.667 and 0.688. By Theorem 4.1, $E_2/U = 1/(1 + \sigma^2/D^2)$, so these two targets have nearly the same

relative dispersion of credit. The sample therefore contains no pair with similar D and U but clearly different concentration; Example 4.2 shows that such pairs are possible. Equal-size subsampling changes the comparison again: P1 has eight distinct authors among its five external works, whereas five works sampled from P3 give about six on average. Neither comparison controls for mathematical topic or citation opportunity.

Table 4: Effect of screening and of retaining identified self-citations.

	Raw external D	Screened external D	All C	All U	All D
P1	0.500	0.625	6	10	0.600
P2	0.697	0.576	24	39	0.615
P3	1.222	1.111	10	9	1.111
P4	1.111	0.889	8	9	0.889
P5	1.150	1.100	30	27	1.111
P6	1.000	0.846	11	13	0.846

The raw external calculation already excludes incomplete author lists and identified self-citations, but is made before consolidation and before removal of incorrect target links; it shows the effect of those two operations and is not an independent estimate.

For P1, OpenAlex linked *Surface bundles and the section conjecture* (work W3092687538) to Roberts's paper. The publisher's reference list instead contains Hillman's *Sections of surface bundles*, in *Geometry & Topology Monographs* 19 (2015), pp. 1–19, and contains no reference to Roberts's target. The source is `https://doi.org/10.1007/s00208-022-02421-9`. Removing this link reduces the raw external count of citing works from six to five and the number of citing authors from twelve to eight, so D changes from 0.500 to 0.625.

For P6, a reference in *Adsorption/desorption of naphthalene and phenanthrene in a binary competitive system in the riparian zone* (work W3210510624) describes a 2009 article on analytical methods for polycyclic aromatic hydrocarbons, but carries the DOI of Schreieder's target paper. The publisher source is `https://doi.org/10.1007/s10653-021-01147-y`. This group also lacks complete author identifiers; by the precedence rule above it is counted as an incorrect link. Since it would have failed the identifier check anyway, its removal does not change the raw external value in Table 4.

For each target, the reference list of one citing work was also inspected, and the target reference was found in all six cases. For P4, the published version contained the reference although an earlier preprint did not. For P5, the checked citing paper is a self-citation and contributes only to the all-citations subset. Six checks are too few to estimate a rate of erroneous reference links.

10.4 Limits of the sample

The pilot uses retrieved and screened records for six papers from one journal and publication year. It is not a random sample of mathematicians or a complete citation census. The nine groups excluded for incomplete identifiers, possible remaining duplicate versions, unresolved author identities, and the count discrepancy for P3 all limit interpretation, and the exclusions may affect different targets differently. Only six reference links were checked by hand, and subject differences remain even within one journal.

The pilot shows that counting and screening decisions can change D substantially in this sample. It does not provide a threshold for unusual citation behaviour, a comparison of research quality, or a general error rate. The supplementary archive preserves the raw responses, manual decisions, group membership, calculations, and reference checks, so that these results can be recomputed or revised.

11 Interpretation

For a specified collection of citation links, U counts the distinct citing authors recorded, R measures the average repetition of citation incidences, and E_2 is the effective number of authors for the fractional distribution. The identity $D = R/\overline{k}$ shows that D must be read together with citing-team size, and the identity (8) shows that mean credit and concentration answer different questions.

The number of distinct citing authors says nothing about whether they belong to one laboratory, several countries, or several disciplines; measuring disciplinary or geographic diversity requires further labels and a rule for authors or papers in more than one category. Repeated citation may reflect continuing use of a theorem, method, or data set, and a small value of D can come from one large team. Counts of people and references cannot by themselves reveal why people cite.

The ratio C/U is therefore one descriptive component of a citation profile. The mathematical results describe how it depends on counting conventions, overlap between citing teams, the amount of material observed, and identification errors. The pilot records concrete changes caused by version consolidation and reference errors; its small, fixed sample does not support rankings or conclusions about whole fields.

Statements and Declarations

Data and code availability. The supplementary archive contains the saved OpenAlex responses, the selection protocol, record-level decisions, analysis code, and calculated tables. The calculations use only Python's standard library and can be reproduced without a paid bibliographic subscription. The spreadsheet in the archive records the calculated values at the time of submission. Full-text articles inspected for reference checks are not redistributed.

Acknowledgment

OpenAI's ChatGPT assisted with drafting and revising the manuscript, examining the mathematical arguments, and preparing code for the preliminary bibliometric analysis. The author is responsible for the final text, mathematical claims, data interpretation, and references.

References

I. Ajiferuke and D. Wolfram, Citer analysis as a measure of research impact: library and information science as a case study, *Scientometrics* **83** (2010), 623–638. doi:10.1007/s11192-009-0127-6.

B. Cappelletti-Montano, S. Columbu, S. Montaldo, and M. Musio, New perspectives in bibliometric indicators: Moving from citations to citing authors, *Journal of Informetrics* **15** (2021), no. 3, 101164. doi:10.1016/j.joi.2021.101164.

L. Egghe, Journal diffusion factors and their mathematical relations with the number of citations and with the impact factor, in P. Ingwersen and B. Larsen (eds.), *Proceedings of ISSI 2005*, 109–120. https://www.issi-society.org/proceedings/issi_2005/Egghe_ISSI2005.pdf.

I. Evdaimon, J. P. A. Ioannidis, G. Nikolentzos, M. Chatzianastasis, G. Panagopoulos, and M. Vazirgiannis, Indicators of small-scale and large-scale citation concentration patterns,

abstract, *10th International Congress on Peer Review and Scientific Publication* (2025). https://peerreviewcongress.org/abstract/indicators-of-small-scale-and-large-scale-citation-concentration-patterns.

F. Franceschini, D. Maisano, A. Perotti, and A. Proto, Analysis of the ch-index: an indicator to evaluate the diffusion of scientific research output by citers, *Scientometrics* **85** (2010), 203–217. doi:10.1007/s11192-010-0165-0.

M. O. Hill, Diversity and evenness: a unifying notation and its consequences, *Ecology* **54** (1973), 427–432. doi:10.2307/1934352.

J. E. Hirsch, An index to quantify an individual's scientific research output, *Proceedings of the National Academy of Sciences USA* **102** (2005), 16569–16572. doi:10.1073/pnas.0507655102.

M. Jovanović and F. Fritsche, There goes another one: Introducing the NUCA-set of indicators, *Proceedings of the 13th International Conference on Scientometrics and Informetrics*, vol. 1 (2011), 333–338. https://www.issi-society.org/proceedings/issi_2011/ISSI_2011_Proceedings_Vol1_36.pdf.

OpenAlex, *Attributes—Works*, documentation, accessed October 29, 2026. https://help.openalex.org/data/works/attributes/.

OpenAlex, *Authorships*, documentation, accessed October 29, 2026. https://help.openalex.org/data/authorships/.

OpenAlex, *Disambiguation—Authors*, documentation, accessed October 29, 2026. https://help.openalex.org/data/authors/disambiguation/.

I. Rowlands, Journal diffusion factors: a new approach to measuring research influence, *Aslib Proceedings* **54** (2002), no. 2, 77–84. doi:10.1108/00012530210435211.