



Quantifying the higher-order influence of scientific publications

Massimo Franceschet¹ · Giovanni Colavizza²

Received: 30 November 2019 / Published online: 13 July 2020
© The Author(s) 2020

Abstract

Citation impact is commonly assessed using direct, first-order citation relations. We consider here instead the indirect influence of publications on new publications via citations. We present a novel method to quantify the higher-order citation influence of publications, considering both direct, or first-order, and indirect, or higher-order citations. In particular, we are interested in higher-order citation influence at the level of disciplines. We apply this method to the whole Web of Science data at the level of disciplines. We find that a significant amount of influence—42%—stems from higher-order citations. Furthermore, we show that higher-order citation influence is helpful to quantify and visualize citation flows among disciplines, and to assess their degree of interdisciplinarity.

Keywords Impact · Influence · Citation impact · Citation influence · Citation networks · Interdisciplinarity

Introduction

New knowledge builds on previous knowledge: this is a central tenet of science. A publication relies on previous publications and cites them to acknowledge this debt (Merton 1957). Although citations acknowledge direct influences, the extent of the influence of a publication can go beyond these *first-order* relations. The study of the influence of previous publications on new ones rests at the core of scientometrics. The visualization and quantification of such dependence has been termed “algorithmic historiography” by Garfield et al. (1964, 2003). A variety of tools have been developed for the purpose of facilitating such exploration (Chen 2006; van Eck and Waltman 2010, 2014; Marx et al. 2014; Thor et al. 2016). Furthermore, previous literature has investigated methods to trace the historical development of science using citations (Lucio-Arias and Leydesdorff 2008; Yi-Ning and Hsu 2016; Subelj et al. 2020) and text (Gerow et al. 2018; Jurgens et al. 2018; Soni

✉ Giovanni Colavizza
g.colavizza@uva.nl

Massimo Franceschet
massimo.franceschet@uniud.it

¹ University of Udine, Via delle Scienze 206, 33100 Udine, Italy

² University of Amsterdam, Postbus 94550, 1090 GN Amsterdam, The Netherlands

et al. 2019). Our related goal here is to quantify citation influence, and thus give credit, beyond direct citations. In particular, we aim at understanding the interplay of first and higher-order influence across academic disciplines.

In this contribution we define *higher-order citations* as citations chains of arbitrary length among pairs of publications, and show how the higher-order citation matrix among disciplines can be computed in an iterative and efficient way. Our proposed method is related to the well-known PageRank algorithm (Brin and Page 1998; Franceschet 2011; Waltman and Yan 2014), but it is specifically focused on quantifying higher-order citation influence. We apply this novel definition to the Web of Science dataset between years 2000 and 2016 included (17,932,523 publications and 190,550,206 citations among them). We show that the contribution of first-order (length 1) citations accounts for 58% of the whole higher-order citation flow, hence it misses a conspicuous part (42%) of citation information. Indeed, higher-order citations bring a clear picture of the relationships among disciplines (Klavans and Boyack 2009). Furthermore, we observe this added value by clustering disciplines into larger communities, finding disciplines that act as brokers among communities, and distinguishing between interdisciplinary and autarchic disciplines.

Methodology

Let $G = (V, E)$ be a citation network with n nodes V and m directed edges E . We assume the nodes represent publications. If publication i cites publication j , then $(i, j) \in E$. Normally, G is a Directed Acyclic Graph (DAG), because citations only go from more recent publications to older publications.¹ A simple example is depicted in Fig. 1.

Let A be the adjacency matrix, so that $A_{ij} = 1$ whenever i cites j , that is $(i, j) \in E$, and $A_{ij} = 0$ otherwise. Let d_i be the outdegree of node i , i.e., the number of publications referenced by publication i within the citation network G .

We then recursively define the dependence of publication i on publication j as the mean dependence of publications referenced by i on publication j :

$$P_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j \text{ and } d_i = 0, \\ \frac{1}{d_i} \sum_k A_{ik} P_{kj} & \text{if } i \neq j \text{ and } d_i > 0. \end{cases}$$

We say that P_{ij} is the *dependence* of i on j , but on the same note it is the *influence* of j on i . Notice that the recursive equation has always a solution since recursion proceeds from each publication to its citing publications, and the graph G is acyclic.

Let us label each edge of the graph (i, j) with probability $1/d_i$ of going from i to j in a random walk on the graph. Given a path $\pi = k_1, k_2, \dots, k_r$ on the graph, we define the *likelihood* of the path π as

$$p(\pi) = \prod_{i=1}^{r-1} \frac{1}{k_i}.$$

¹ There are some exceptions, but these can be removed so as to ensure that G is a DAG.

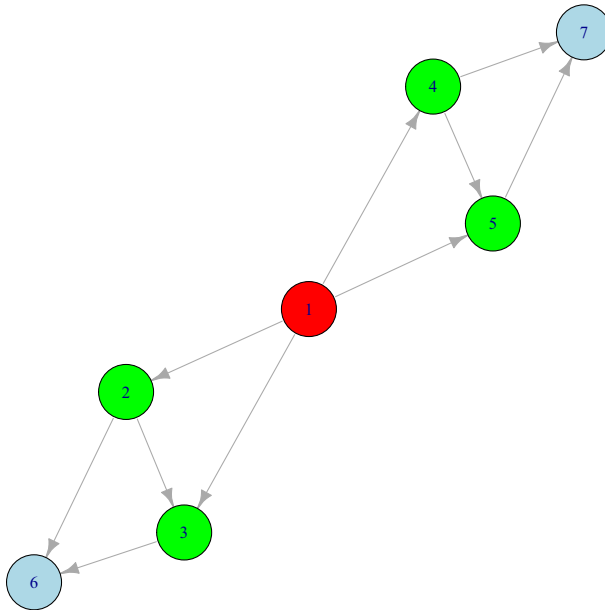


Fig. 1 A publication citation network. The red node is the root (a node with no predecessors) of the DAG, the blue nodes are the leaves (nodes with no successors) and the green nodes are intermediate nodes (vertices with both predecessors and successors). (Color figure online)

The dependence P_{ij} , when $i \neq j$, is then the sum of likelihoods of all paths from i to j in the graph. In general:

The dependence P_{ij} is large if there are numerous likely paths starting at i and ending in j

For instance, with reference to the graph in Fig. 1, we have:

$$\begin{aligned}
 P_{11} &= 1 \\
 P_{12} = P_{14} &= \frac{1}{4} \\
 P_{13} = P_{15} &= \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{1}{1} = \frac{3}{8} \\
 P_{16} = P_{17} &= \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{1}{2} \cdot \frac{1}{1} + \frac{1}{4} \cdot \frac{1}{1} = \frac{1}{2}
 \end{aligned}$$

We can write this more compactly using matrix notation. Let D be a diagonal matrix such that $D_{ii} = \frac{1}{d_i}$ if $d_i > 0$ and $D_{ii} = 0$ otherwise. We can then write

$$P = DAP + I \tag{1}$$

where I is the $n \times n$ identity matrix. We can solve for P and obtain

$$P = (I - DA)^{-1}$$

Notice that, if we topologically sort the nodes in A (as done in Fig. 1), which is possible since G is a DAG, then both A and $I - DA$ are triangular matrices. In particular, the diagonal elements of $I - DA$ are equal to 1. Hence $\det(I - DA) = 1$, the matrix $I - DA$ is

invertible and Eq. (1) has a solution, as noticed above. The inverse $P = (I - DA)^{-1}$ is also triangular.

One can also iteratively compute P using the fact that:

$$P = \sum_{i=0}^{\infty} (DA)^i = \sum_{i=0}^l (DA)^i \tag{2}$$

where $l \leq n - 1$ is the longest path in the graph G and n is the number of nodes of G . The last equality holds because G is acyclic and thus $(DA)^i = 0$ for all $i > l$. We expect $l \ll n$. In particular, the length l is bounded by the longest path in the dataset, which corresponds to the number of time instants in the granularity of the dataset. For instance, if the dataset covers 10 years and publication dates are given with a month granularity, then l is lower than $12 \cdot 10 = 120$.

Matrix $(DA)^i$ computes the dependence contribution of paths of length i in graph G . In particular, for $i = 1$, the matrix DA represents *first-order citations*, that is direct citations among publications. On the other hand, matrix $(DA)^i$ for $i > 1$, encodes *higher-order citations*, that is chains of citations of length i among publications.

Notice that $P_{ij} \neq 0$ if and only if there exists at least on path from i to j in graph G . Hence, matrix P has the same non-zero pattern of the adjacency matrix of the transitive closure of G . We thus expect P to be denser than A .

Discipline dependence

Instead of looking at the individual dependence of publication i on publication j , we are interested in disciplinary dependencies. In particular, we are interested in the dependence of a publication (or of a discipline) on a discipline.

Let us denote by Q_{iv} the extent to which publication i belongs to discipline v , hence Q is a matrix $n \times k$, where n is the number of publications and k is the number of disciplines. For the non-overlapping case, $Q_{iv} = 1$ if publication i belongs to discipline v . A publication can belong to multiple disciplines, thus $Q_{iv} > 0$ for possibly more than a single discipline v . In either case, we have $\sum_v Q_{iv} = 1$ and $Q_{iv} \geq 0$.

The dependence R_{iv} of publication i on discipline v can then be defined as the sum of the dependencies of publication i on articles in v :

$$R_{iv} = \sum_j P_{ij} Q_{jv},$$

or, in matrix notation

$$R = PQ.$$

Note that

$$\begin{aligned} R &= PQ \\ &= (DAP + I)Q \\ &= DAPQ + Q \\ &= DAR + Q. \end{aligned}$$

We can hence iteratively compute matrix R without materializing matrix P :

$$\begin{cases} R^{(0)} &= Q \\ R^{(i+1)} &= DAR^{(i)} + Q \end{cases}$$

Notice that $R^{(i)} = \sum_{j=0}^i (DA)^j Q$ is the dependence contribution of citation paths of length up to i . Hence

$$R = \sum_{i=0}^{\infty} (DA)^i Q = \sum_{i=0}^l (DA)^i Q$$

where l is the longest path in the graph, and the iterative computation of R can stop after l steps. Although R can be as dense as P , it has size $n \times k$, which is more manageable than the size of P , which is $n \times n$, since we expect $k \ll n$.

As a particular case, the dependence r_i of publication i on the whole network is $r_i = \sum_j P_{i,j}$, that is, $r = Pe$. We thus have that:

$$r = Pe = (DAP + I)e = DAPE + e = DAR + e.$$

Recall that the Pagerank of G , with damping factor α and exogenous vector β , is the vector x such that $x = \alpha DAx + \beta$ (Newman 2018). Hence, interestingly, the dependence vector r is also the Pagerank of G with damping factor $\alpha = 1$ and exogenous vector $\beta = e$.

One can also define the the dependence $S_{u,j}$ of discipline u on publication j as the sum of the dependence of publications in u on article j :

$$S_{uj} = \sum_i Q_{iu} P_{ij},$$

or, in matrix notation

$$S = Q^T P.$$

Notice that since $P = (I - DA)^{-1}$, then $P(I - DA) = I$ and hence $P = PDA + I$. It follows that $S = SDA + Q^T$ and also S can be computed iteratively.

The dependence F_{uv} of discipline u on discipline v is the sum of the dependence of papers in u on papers in v , that is:

$$F_{uv} = \sum_i Q_{iu} R_{iv} = \sum_i \sum_j Q_{iu} P_{ij} Q_{jv},$$

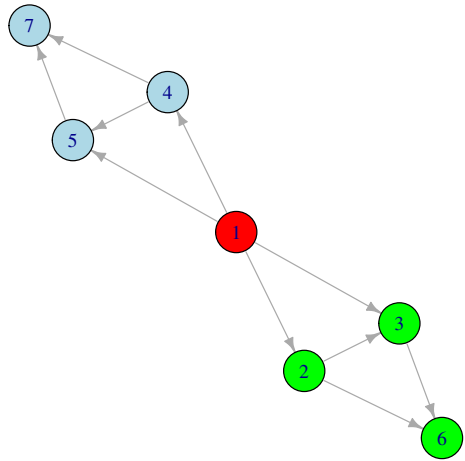
or, in matrix notation

$$F = Q^T R = Q^T P Q = S Q.$$

We also define $F^{(i)} = Q^T R^{(i)}$, for $i \geq 0$, as the citation flow matrix for paths of length up to i . Notice that, for $i \geq 1$, $F^{(i)} - F^{(i-1)}$ is the citation flow matrix for paths of length equal to i .

Consider again the simple citation network depicted in Fig. 2, where nodes are partitioned in 3 disjoint disciplines. The light blue and green communities are closed worlds (autarchies), since they reference only within their own groups (their off-diagonal flows in matrix F is indeed 0). On the other hand, the red community is more interdisciplinary, since it references the other two groups outside its territory (the off-diagonal flow in matrix F is 2.25).

Fig. 2 A publication citation network where nodes are partitioned in 3 non-overlapping disciplines



Case study

We applied our method on all publications from the CWTS in-house version of the Web of Science, considering the years between 2000 and 2016 included. We consider a total of 17,932,523 publications, and 190,550,206 citations among them—excluding 444,436 synchronous citations, which we discarded to guarantee that G is a DAG.² The longest citation path in the dataset is of length 29—equal to the maximum number of iterations needed for convergence. In what follows, we rely on the high-level aggregation of the journal-based classification of Web of Science, which represents 30 broad disciplines (see Table 2).

The contribution of higher-order citations

We start by assessing the contribution of first-order and higher-order citations to the citation flow among disciplines. Recall that partial flow matrix $F^{(i)}$ is the flow matrix for paths of length up to i , with total flow matrix $F = F^{(l)}$, where l is the length of the longest path in the citation graph. Let $M^{(i)} = F^{(i)} - F^{(i-1)}$ be the flow matrix for paths of length precisely i . The entry-wise matrix norm $\|\cdot\|_1$ defined as $\|M^{(i)}\|_1 = \sum_{u,v} |M_{u,v}^{(i)}|$ is a measure of the total citation flow contained in matrix $M^{(i)}$. We also tested the Frobenius norm $\|\cdot\|_2$ with similar outcomes.

We computed the norm of partial flow matrices $M^{(i)}$ relative to the norm of total flow matrix $F = F^{(l)}$, for $1 \leq i \leq l$. Results are shown in Fig. 3. First-order (direct) citations contribute for 58% to the overall flow, hence higher-order citations contribute for 42%, a significant share. In particular, the share of second-order (length 2) citations is 20%, that of third-order citations (length 3) is 12%, and that of fourth-order citations (length 4) is 6%. Longer citations paths account for about 4% of the flow. When we consider the top disciplines by flow contribution (Fig. 4), we have that six of them account for 38% (over 42%) of first-order flow, 13% (over 20%) of second-order flow, 8% (over 12%) of third-order flow, and 4% (over 4%) of fourth-order flow, following a similar pattern to global

² A citation between two publications is discarded if the publication time (year and month) of the citing publication is the same, or older than the publication time of the cited publication.

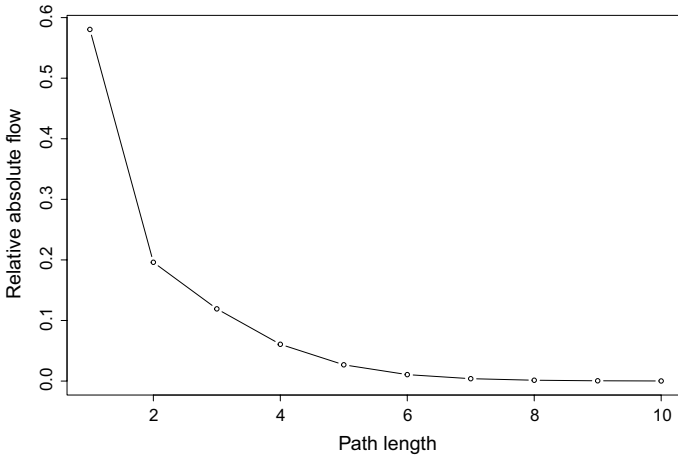


Fig. 3 Relative contribution to the flow of citation paths at given orders (path lengths)

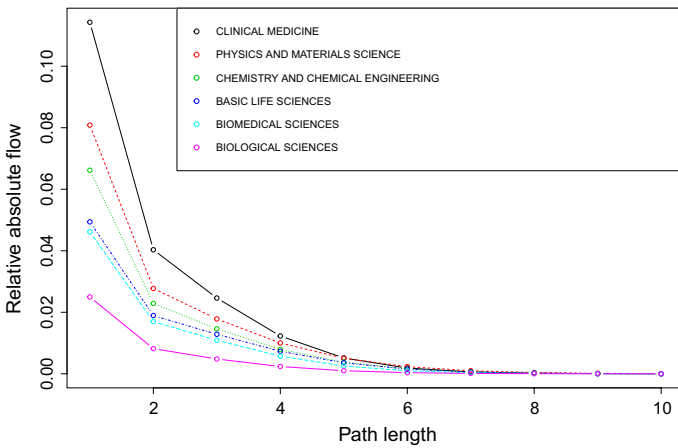


Fig. 4 Relative contribution to the flow of citation paths at given orders (path lengths), for the top 6 disciplines by relative flow contribution

contributions.³ We conclude that there is an important part of dependence flow that goes beyond direct citations which is worth investigating.

³ In order: Clinical medicine, Physics and materials science, Chemistry and chemical engineering, Basic life sciences, Biomedical sciences, Biological sciences.

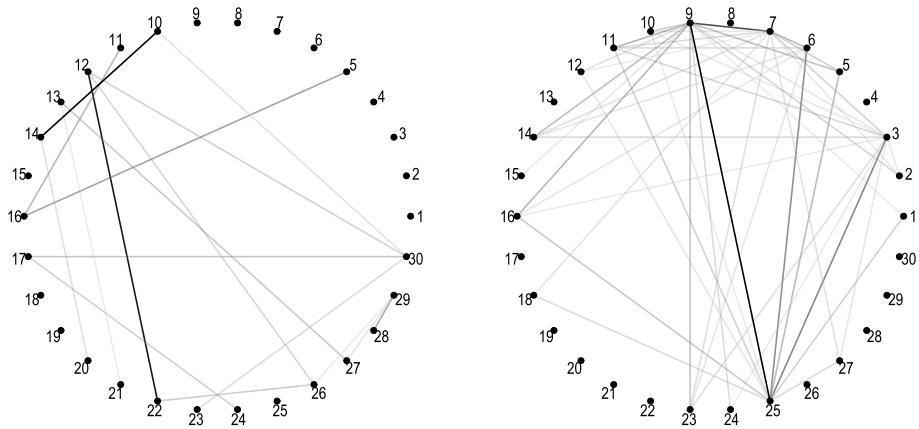


Fig. 5 The higher than expected flows (left) and lower than expected flows (right) among disciplines. E.g., Computer Sciences (10) and Electrical Engineering and Telecommunication (14), as well as Economics and Business (12) and Management and Planning (22) reference each other more than expected, while disciplines Clinical Medicine (9) and Physics and Materials Science (25) reference each other less than expected. See Table 2 for the names of the disciplines

The citation flow network

The citation flow matrix is a full matrix and hence the corresponding flow network is a full graph. However, one might investigate the pairs of disciplines that have an higher than expected citation flow, and those that have a lower than expected citation flow.

Table 2 contains, for each discipline, the internal citation flow (self-flow), the outgoing and incoming citation flows and, moreover, the size of the discipline in number of articles. As expected, citation flows are strongly correlated with size of the discipline (Pearson correlation above 0.9).

To overcome the size-dependence issue, we normalize the flow matrix using the signed contribution to Pearson’s χ -squared test. The normalized flow $\hat{F}_{i,j}$ between disciplines i and j is computed as:

$$\hat{F}_{i,j} = \frac{F_{i,j} - E_{i,j}}{\sqrt{E_{i,j}}}$$

where

$$E_{i,j} = \frac{(\sum_k F_{i,k}) \cdot (\sum_k F_{k,j})}{\sum_{u,v} F_{u,v}}$$

is the *expected* flow between i and j . The pairs of disciplines that significantly cite each other more than expected (above the 90th percentile) and less than expected (below the 10th percentile) are shown in Fig. 5. As for within-discipline citation flows (normalized by expected citations), Astronomy and Astrophysics, Mathematics, and Language and Linguistics lead the ranking, while Instruments and Instrumentation, Basic Medical Sciences and General and Industrial Engineering are at the bottom.

Furthermore, we consider the same network limited to positively weighted edges, thus with a higher than expected citation flow. We then apply the fast greedy clustering method to this

Fig. 6 The higher-order citation flow network limited to positive (more likely) edges, and divided into communities. Cyan: life and medical sciences; Purple: Earth and environment sciences; Red: mathematical sciences; Green: social and human sciences. We highlight disciplines with large betweenness centrality: Environmental Sciences and Technology (16), Health Sciences (18), and General and Industrial Engineering (17) lead the ranking. See Table 2 for the names of the disciplines. Compare with first-order graph in Fig. 7. (Color figure online)

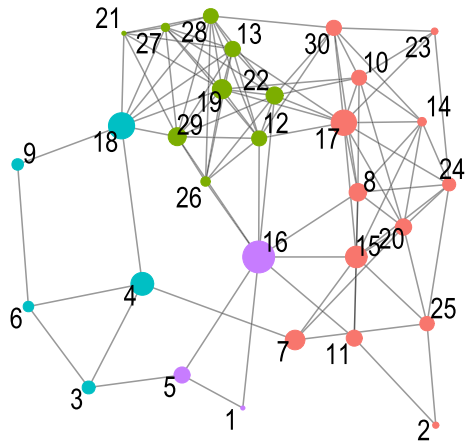
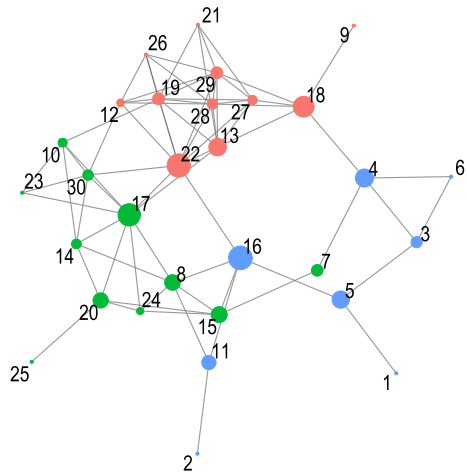


Fig. 7 The first-order citation flow network limited to positive (more likely) edges, and divided into communities. See Table 2 for the names of the disciplines. Compare with higher-order graph in Fig. 6



network, as depicted in Fig. 6. Four macro areas emerge from this analysis, namely the life and medical sciences, science and engineering applied to the Earth and the environment, mathematical sciences and social and human sciences. If we do the same limiting ourselves to first-order citations (Fig. 7), the partition of disciplines into communities is less clear.

Our analyses suggest that some disciplines are more interdisciplinary (connecting different communities) and other more autarchic (mostly self-referencing), a topic we explore in the following section.

Interdisciplinarity and autarchy

In this section we match higher-order citation flows with measures of interdisciplinarity. We claim that:

A discipline is interdisciplinary when it is evenly cited from dissimilar disciplines.

This thesis immediately recalls the *Rao quadratic entropy* (Rao 1982), which has been previously used to measure interdisciplinarity (Porter and Rafols 2009; Rafols and Meyer

2010; Yegros-Yegros et al. 2015; Wang and Schneider 2019). The Rao quadratic entropy is one measure among others which have been studied in the literature (Mugabushaka et al. May 2016). Let us consider a set of objects and a probability distribution p such that p_i is the probability of object i . Suppose we also have information about pairwise distance (dissimilarity) d_{ij} among any two objects i and j . Then a measure of heterogeneity among objects is the Rao quadratic entropy:

$$R(p, d) = \sum_{i,j} p_i p_j d_{i,j}$$

There are two components in this definition of heterogeneity: (1) the evenness of the distribution p , (2) the distances d among objects. It holds that, in general:

- $R(p, d)$ is large when p evenly distributes its probability among dissimilar objects;
- on the contrary, $R(p, d)$ is small when p concentrates its probability on similar objects.

To apply Rao's measure to the higher-order citation flow matrix F , we proceed as follows. For each discipline pair u and v , let

$$p_{u,v} = \frac{F_{u,v}}{\sum_i F_{i,v}}$$

Notice that $p_{u,v}$ is the relative share of citation flow from discipline u to discipline v compared to the total flow received by v . Notice, moreover, that $p_{*,v} = (p_{1,v}, p_{2,v}, \dots, p_{k,v})$ is a probability distribution.

The similarity $s_{u,v}$ among two disciplines u and v is computed as the cosine of the angle between the u and v columns $F_{*,u}$ and $F_{*,v}$ of the flow matrix F :

$$s_{u,v} = \cos(F_{*,u}, F_{*,v}) = \frac{F_{*,u} F_{*,v}}{\|F_{*,u}\| \|F_{*,v}\|}$$

The cosine runs from 0 (no similarity) to 1 (maximum similarity). Hence, two disciplines are similar if they have a similar pattern of incoming citation flows. The distance $d_{u,v}$ among two disciplines u and v is then

$$d_{u,v} = 1 - s_{u,v}$$

so that two disciplines are distant if they are not similar.

Finally, for each discipline v , we apply the Rao quadratic entropy to the flow distribution $p_{*,v}$ and distance measure d among disciplines. This gives us a measure of interdisciplinarity for each discipline. The top and bottom 5 interdisciplinary disciplines are given in Table 1.

Notice how two interrelated disciplines like Statistical Sciences and Mathematics end up on quite different ranks: while Statistics is interdisciplinary, Mathematics is rather autarchic. Indeed, Mathematics receives 78% of higher-order citation flow from itself, and the rest from a small number of other fields, mainly Physics, Materials Science and Computer Science. On the other hand, the internal flow for Statistics is limited to 43%. Statistics receives instead a significant citation flow from many other disciplines, including Mathematics, Computer Sciences, Economics and Business, General and Industrial Engineering, Electrical Engineering and Telecommunication, Clinical Medicine. This suggests that higher-order citations should be considered when assessing the degree of interdisciplinarity or autarchy of a discipline.

Table 1 Top 5 (top) and bottom 5 (bottom) disciplines by their interdisciplinarity

Discipline	Rao
Statistical sciences	0.678
Management and planning	0.645
General and industrial engineering	0.641
Social and behavioral sciences, interdisciplinary	0.622
Civil engineering and construction	0.601
...	...
Chemistry and chemical engineering	0.360
Mathematics	0.341
Astronomy and astrophysics	0.316
Physics and materials science	0.302
Clinical medicine	0.294

Conclusion

A considerable amount of effort goes into quantifying and assessing citation influence and impact via direct citations. We proposed instead here to quantify citation influence beyond direct citations by also using *higher-order citations*, that is citations chains of arbitrary length among pairs of publications. We have presented a method, informed by PageRank, to quantify the higher-order citation influence of publications. The proposed method accounts for both direct, or first-order, and indirect, or higher-order citations. In particular, we assessed the method on the whole Web of Science corpus between 2000 and 2016 at the level of entire disciplines.

Our results show that the contribution of first-order (length 1) citations accounts for 58% of the whole higher-order citation flow, while higher-order citations (levels 2 and above) account for 42%: a significant share. The proposed method is size-dependent, yet easily normalized, and it can be used for a variety of applications. We investigated two here. By using higher-order citation flows, we were able to provide for a high-level map of science clearly distinguishing among four macro-areas: life and medical sciences, Earth and environment sciences, mathematical sciences, social and human sciences. The same picture using only first-order information was found to be less clear-cut. Furthermore, we used the proposed method to rate disciplines according to their degree of interdisciplinarity using the Rao quadratic entropy. We are thus able to distinguish between autarchic disciplines, e.g., mathematics, and interdisciplinary ones, e.g. statistics. We suggest that accounting for higher-order citations is thus relevant and important, and might help on a variety of open scientometrics questions: performing clustering, measuring interdisciplinarity, assessing the impact of fundamental research, among others.

Acknowledgements This work stems from prior efforts in collaboration with Ludo Waltman and Vincent A. Traag Colavizza et al. (2019), whom we thank for their contribution. We are grateful to the Centre for Science and Technology Studies (CWTS), Leiden University, for providing us access to their databases.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the

material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Appendix

See Table 2.

Table 2 The Web of Science disciplines, with fields id, name of discipline, size, self citation flow, incoming citation flow and outgoing citation flow

id	Discipline	Size	Self flow	Incoming flow	Outgoing flow
1	Agriculture and food science	875,440.50	780,500.12	529,167.52	743,893.92
2	Astronomy and astrophysics	381,254.75	686,101.56	219,418.90	171,588.10
3	Basic life sciences	2,579,591.25	3,456,087.00	3,474,212.42	2,007,738.04
4	Basic medical sciences	268,307.25	199,883.83	335,008.55	483,618.99
5	Biological sciences	1,402,123.00	1,259,296.75	910,008.50	1,164,499.91
6	Biomedical sciences	2,507,916.50	2,356,196.00	2,470,855.73	2,487,821.15
7	Chemistry and chemical engineering	3,510,294.25	4,352,712.50	1,959,569.08	2,466,840.06
8	Civil engineering and construction	160,902.86	127,872.16	132,699.23	155,468.25
9	Clinical medicine	6,024,741.50	8,482,322.00	3,270,959.40	3,051,526.10
10	Computer sciences	647,474.88	668,669.81	482,215.10	506,644.46
11	Earth sciences and technology	934,568.50	1,395,625.38	549,727.39	443,447.22
12	Economics and business	429,852.88	526,190.56	277,452.46	185,736.94
13	Educational sciences	238,509.97	212,864.89	116,494.86	163,714.45
14	Electrical engineering and telecommunication	842,418.88	902,059.25	629,718.60	612,375.60
15	Energy science and technology	343,416.62	196,160.98	263,133.26	337,039.66
16	Environmental sciences and technology	983,358.88	1,125,205.62	886,273.78	1,027,153.47
17	General and industrial engineering	198,930.95	101,423.06	163,249.88	222,303.95
18	Health sciences	496,159.94	479,285.53	429,532.91	612,249.92
19	Information and communication sciences	104,181.30	79,418.53	56,385.11	76,125.22
20	Instruments and instrumentation	154,830.81	59,613.47	153,544.22	185,356.41
21	Language and linguistics	98,703.09	80,662.09	24,108.05	42,272.65
22	Management and planning	156,367.38	115,467.25	143,213.02	145,707.05
23	Mathematics	831,350.88	1,003,179.06	281,315.60	334,351.78
24	Mechanical engineering and aerospace	595,979.12	489,624.09	386,884.15	441,418.25
25	Physics and materials science	4,089,318.25	6,163,358.50	20,98,397.77	2,250,967.82
26	Political science and public administration	193,848.67	170,155.39	83,619.01	76,208.32
27	Psychology	581,770.75	617,750.44	458,871.44	412,155.69
28	Social and behavioral sciences, interdisciplinary	132,240.47	74,401.98	108,292.33	128,959.83
29	Sociology and anthropology	218,277.44	172,026.30	148,015.28	173,080.76
30	Statistical sciences	222,210.95	194,457.47	252,692.38	184,771.96

Note that the size is the sum of article classifications by discipline. An article can belong to multiple disciplines

References

- Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7), 107–117.
- Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3), 359–377.
- Colavizza, G., Franceschet, M., Traag, V. A., & Waltman, L. (2019). Quantifying the long-term influence of scientific publications. In *Proceedings of the 17th international conference on scientometrics & informetrics*.
- Franceschet, M. (2011). PageRank: Standing on the shoulders of giants. *Communications of the ACM*, 54(6), 92–101.
- Garfield, E., Pudovkin, A. I., & Istomin, V. S. (2003). Why do we need algorithmic historiography? *Journal of the American Society for Information Science and Technology*, 54(5), 400–412.
- Garfield, E., Sher, I. H., & Torpie, R. J. (1964). *The use of citation data in writing the history of science* (Vol. 49, No. 638, p. 1256). The Institute for Scientific Information, Technical Report, AF.
- Gerow, A., Hu, Y., Boyd-Graber, J., Blei, D. M., & Evans, J. A. (2018). Measuring discursive influence across scholarship. *Proceedings of the National Academy of Sciences*, 115, 3308–3313.
- Jurgens, D., Kumar, S., Hoover, R., McFarland, D., & Jurafsky, D. (2018). Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6, 391–406.
- Klavans, R., & Boyack, K. W. (2009). Toward a consensus map of science. *Journal of the American Society for Information Science and Technology*, 60(3), 455–476.
- Lucio-Arias, D., & Leydesdorff, L. (2008). Main-path analysis and path-dependent transitions in HistCite TM-based historiograms. *Journal of the American Society for Information Science and Technology*, 59(12), 1948–1962.
- Marx, W., Bornmann, L., Barth, A., & Leydesdorff, L. (2014). Detecting the historical roots of research fields by reference publication year spectroscopy (RPYS): Detecting the Historical Roots of Research Fields by Reference Publication Year Spectroscopy (RPYS). *Journal of the Association for Information Science and Technology*, 65(4), 751–764.
- Merton, R. K. (1957). Priorities in scientific discovery: A chapter in the sociology of science. *American Sociological Review*, 22(6), 635–659.
- Mugabushaka, A.-M., Kyriakou, A., & Papazoglou, T. (2016). Bibliometric indicators of interdisciplinarity: The potential of the Leinster–Cobbold diversity indices to study disciplinary diversity. *Scientometrics*, 107(2), 593–607.
- Newman, M. E. J. (2018). *Networks: An introduction* (2nd ed.). Oxford: Oxford University Press.
- Porter, A. L., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, 81(3), 719–745.
- Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, 82(2), 263–287.
- Rao, C. R. (1982). Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, 21, 24–43.
- Soni, S., Lerman, K., & Eisenstein, J. (2019). Follow the leader: Documents on the leading edge of semantic change get more citations. arXiv:1909.04189 [physics]
- Subelj, L., Waltman, L., Traag, V., & van Eck, N. J. (2020). Intermediacy of publications. *Royal Society Open Science*, 7(1), 190207.
- Thor, A., Marx, W., Leydesdorff, L., & Bornmann, L. (2016). Introducing CitedReferencesExplorer (CRExplorer): A program for reference publication year spectroscopy with cited references standardization. *Journal of Informetrics*, 10(2), 503–515.
- van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523–538.
- van Eck, N. J., & Waltman, L. (2014). CitNetExplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics*, 8(4), 802–823.
- Waltman, L., & Yan, E. (2014). PageRank-related methods for analyzing citation networks. In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), *Measuring scholarly impact* (pp. 83–100). Berlin: Springer.
- Wang, Q., & Schneider, J. W. (2019). Consistency and validity of interdisciplinarity measures. *Quantitative Science Studies*, 1, 239–263.
- Yegros-Yegros, A., Rafols, I., & D’Este, P. (2015). Does interdisciplinary research lead to higher citation impact? The different effect of proximal and distal interdisciplinarity. *PLOS ONE*, 10(8), e0135095.
- Yi-Ning, T., & Hsu, S.-L. (2016). Constructing conceptual trajectory maps to trace the development of research fields. *Journal of the Association for Information Science and Technology*, 67(8), 2016–2031.