Evolution of interdependent co-authorship and citation networks

Chakresh Kumar Singh, Demival Vasques Filho, Shivakumar Jolad & Dion R. J. O'Neale

ONLINE
FIRST

Springer

Springer

Check for updates

Evolution of interdependent co-authorship and citation networks

Chakresh Kumar Singh[1] · Demival Vasques Filho[2,3] · Shivakumar Jolad[4] · Dion R. J. O'Neale[2]

Abstract

Studies of bibliographic data suggest a strong correlation between the growth of citation networks and their corresponding co-authorship networks. We explore the interdependence between evolving citation and co-authorship networks focused on the publications, by Indian authors, in American Physical Society journals between 1970 and 2013. We record interactions between each possible pair of authors in two ways: first, by tracing the change in citations, they exchanged and, second, by tracing the shortest path between authors in the co-authorship network. We create these data for every year of the period of our analysis. We use probability methods to quantify the correlation between citations and shortest paths, and the effect on the dynamics of the citation-co-authorship system. We find that author pairs who have a co-authorship distance $d \leq 3$ significantly affect each other's citations, but that this effect falls off rapidly for longer distances in the co-authorship network. The exchange of citation between pairs with $d = 1$ exhibits a sudden increase at the time of first co-authorship events and decays thereafter, indicating an ageing effect in collaboration. This behaviour suggests that the dynamics of the co-authorship network appear to be driving those of the citation network rather than vice versa. Moreover, the majority of citations received by most authors are due to reciprocal citations from current, or past, co-authors. We conclude that, in order to answer questions on the nature and dynamics of scientific collaboration, it is necessary to study both co-authorship and citation network simultaneously.

Keywords Co-authorship networks · Citation networks · Evolution · Network structure

Introduction

Development of scientific theories and technology is a result of continuous interaction, creation, and effective diffusion of ideas between researchers in the knowledge ecosystem. Digitization of publications and advancements in communication technology have made it easier for researchers to be aware of the existing knowledge capital and possible gaps in

✉ Chakresh Kumar Singh
chakresh.singh@iitgn.ac.in

Extended author information available on the last page of the article

the field of study. This facilitation of the spread of scientific knowledge helps researchers to refine their research methods and to contextualize their work within the domain. It also establishes an indirect interaction between individuals. One might not know a researcher personally but is still aware of her work through technical literature and can gain a sense of familiarity with it. Researchers attend gatherings and conferences to broaden their scope of a subject area and look for new ideas and open problems. Awareness to the state of the art and motivation to solve open problems often becomes a factor in setting up new collaborations between individuals.

Publication of scientific articles provides a narrow, but a well-quantified record of collaboration and exchange of technical information. Interactions between researchers can either be by citing one another or by co-authoring papers together, resulting in a complex system that is changing over time. With such a complex and dynamic system at hand, it is interesting to look for any possible underlying pattern hidden in interactions between researchers, and whether these interactions have any mathematical structure. Such a structure can be used to explain the spread of knowledge and the growth of research fields.

Tools and methods developed within the framework of network science have proven to be effective in addressing questions of such nature both quantitatively and qualitatively (Newman 2001a, b; Sinatra et al. 2016). The study of complex systems by using networks, combined with easy access to vast databases of publications have attracted much attention in the last two decades. It has also lead to much research on the structure and evolution of scientific collaboration (Dong et al. 2017; Tomassini and Luthi 2007).

Detailed publication records make it easier to create citation networks and co-authorship networks. While the former are directed networks where edges represent citations between nodes, the latter are undirected (either weighted or unweighted) networks where edges exist between two nodes sharing authorship on a paper. Nodes in these networks can either be papers, authors, universities and so on, depending on the research question of interest.

Individual relationships between authors have a high impact on interactions between institutions at meso-level, and countries at macro-level (Tomassini and Luthi 2007), shaping the changing trend of collaboration. Over time, the pattern of collaboration has shown a shift from individual efforts to more cooperative research, increasing the productivity and diversity of scientific publications globally, and resulting in an increase of innovation in this century (Dong et al. 2017). Co-authorship networks have been shown to exhibit small-world characteristics and display high levels of clustering (Newman 2004). As these networks grow and new authors appear, the network structure changes (Huang et al. 2008; Martin et al. 2013; Vasques Filho 2018). Growing networks can also change authors' topological position in the network structure, which is directly related to one's productivity and popularity (Newman 2001b, a, 2004). The mechanism for evolving co-authorship networks has been shown to exhibit an underlying preferential attachment process (Barabási et al. 2002; Chen et al. 2013). Co-authorship networks are known to become more connected over time, indicating an increase in collaboration between authors. As a result, the average node distance in the network shrinks. Knowing the co-authorship network structure and its evolution also makes it possible to predict future links between authors by exploiting the changes in an author's neighbourhood structure (Huang et al. 2008). Studying co-authorship networks can explain the emergence of new research groups, the significance of some lead researchers, and how one's collaborators change over time.

Citation networks, on the other hand, capture the patterns in generation and diffusion of ideas in the scientific community. Citations received by publications play a significant role in determining their impact as well as their authors' significance in the community (Sinatra et al. 2016). Evolution of citing patterns can be correlated with the evolution of research

fields (Gualdi et al. 2011; Shi et al. 2009). Since citation networks are directed; tree-like hierarchical structures form the backbone of citations. Patterns in citation dynamics have been extensively explored and modelled. The critical feature in citation patterns is the presence of a delay before a paper receives initial citations. Citations acquired by a paper typically increase shortly after publication and reach a maximum within the first few years before decaying with time (Pan et al. 2018; Higham et al. 2017). Considering that the ageing effect is essential to quantify the probability of getting cited (Börner et al. 2004) or the strength of collaboration in citations and co-authorship networks, respectively, the effect of ageing can be accounted for in different ways. It can be a weighted measure on edges, that decays over the time from the contributing authors last shared publication (Fiala et al. 2015), proportional to the time difference between simultaneous participation (Tutoky and Paralič 2011). Instead of decaying weights on the edges, adding an ageing effect on nodes in the citation networks can also determine the changing probability of receiving citation (Hajra and Sen 2005).

Co-authorship and citation networks together reflect the structure and growth of scientific collaboration. Every new publication results in co-authorship and citation events; therefore, it is intuitive that citation and co-authorship networks are interrelated and should have a strong positive correlation in their respective evolution. Many studies addressed these networks and pointed out strong interdependent relations between evolving citations and co-authorship networks (Kas et al. 2012; Keegan et al. 2013; Martin et al. 2013; Amblard et al. 2011; Ding 2011; Tol 2011; Glänzel and Thijs 2004). Network measures on time-varying graphs for both citation and co-authorship networks exhibit co-dependence of these networks in citing patterns and formation of communities (Amblard et al. 2011). Topic modelling algorithms used to assign topics to papers and to investigate citation patterns between authors showed close collaboration between authors working on similar topics and that high profile authors do not generally co-author with one another but significantly cite each other (Ding 2011). Large scale dataset studies (Wallace et al. 2012; Martin et al. 2013) solidify the notion of strong interdependence between citation and co-authorship networks. Citation exchange calculated up to a limited depth of co-authorship connections reveal large gaps in citing patterns of co-authors between natural sciences and social sciences and that the rate of self-citations is constant (Wallace et al. 2012). A detailed analysis of citation and co-authorship networks constructed from a large longitudinal dataset (100 years) of publications in Physical Review journals investigated the temporal changes in citing patterns between collaborators (Martin et al. 2013). One of the main findings of the latter was the stable fraction of self-citations and citations among co-authors with a strong tendency towards reciprocal citations.

The existing interdependence between the two networks has also helped to define sophisticated weighted measures to distribute the credit of citations between co-authors of a paper, resulting in a more efficient way to calculate authors' significance (Tol 2011). Studying the citation and co-authorship networks together not only helps in quantifying a researcher's contribution to the field (González-Teruel et al. 2015), but network centrality measures have also proved to be important in quantifying the effect of citation and co-authorship networks on each other (Biscaro and Giupponi 2014). It has been observed that an author's (node's) centrality value in the co-authorship network is a significant factor behind the number of citations received by them (Sarigöl et al. 2014). Considering the effect of co-authorship networks is also important for defining more sophisticated growth mechanisms for citing patterns in networks (Guo et al. 2017). Combined co-authorship and citation networks have been used to predict new collaboration opportunities; that is, new edges in co-authorship networks (Lande and Andrushchenko 2016). Moreover, together they are useful for quantifying the effect of

authors and their affiliated institutes' international collaborations and region on citations received (Sin 2011; Yan and Ding 2012) by them. Studying both the networks together also helps in forming ranking measures for institutes and researchers (Xu et al. 2017; Zhan and Tse 2017) in scientific collaboration.

Simultaneous analyses of citation and co-authorship networks have given insights into correlations between the two types of networks. Considering the effect of one network on the other gives a better understanding of the true nature of scientific collaborations. While earlier studies have addressed citation and co-authorship networks simultaneously and established a strong interdependence between the two, there is still scope to understand the details of the correlation between these networks. In this study, we build on the strong correlation between interacting pairs of authors in citation and co-authorship networks. First, we track the evolution of relationships between each possible author–author pair in both types of network. Next, we formulate a null hypothesis for the probability of citation exchange between a pair of authors and use probabilistic analysis to compare it with empirical observations from networks constructed using the publications in the American Physical Society (APS) journals between 1970 and 2013. This way, we capture both macroscopic and microscopic changes in network structure and address several questions which would otherwise be challenging to answer.

1. What fraction of authors exchange citations but do not co-author; that is, are indirectly connected in the co-authorship network?
2. How are citations exchanged between co-authors?
3. How do the statistics in 1 vary with network distance between authors?
4. How does receiving a new citation affect the likelihood of an author creating a new link in the co-authorship or the citation network?
5. What is the relationship between the probability of citations and network distance?
6. What is the waiting time distribution for consecutive co-authorship events and consecutive co-citation events?

Our analysis is based on a similar approach used by earlier studies (Barabási et al. 2002; Biscaro and Giupponi 2014; Chen et al. 2013; Kas et al. 2012; Martin et al. 2013; Yan and Ding 2012). However, our method is significantly different from theirs. By tracing the interactions between all possible pairs of authors in the citation network and for all possible shortest paths in the associated co-authorship network, we can see in detail the effect of collaborations on citations and vice-versa. Our main contribution is to explain the effect of the distance in the co-authorship network on the citations exchanged between pairs of authors.

The remainder of this paper is organized as follows: "Methods" section presents our methods and explains how we create the distance and citation matrices, as well as how we perform our empirical calculations. In "Results and discussions" section, we present and discuss our results. We summarize our findings and possible extension of this work in "Conclusion" section.

Methods

For our study, we constructed a longitudinal dataset of publications by Indian researchers in the American Physical Society (APS) journals between 1970 and 2013. Here we consider an author to be Indian if they have used an Indian institution as their address on any publication. Therefore all papers with authors having at least one Indian affiliation

are included in the dataset. There were 14,703 such papers (Singh and Jolad 2019). For the extracted papers, we performed name disambiguation on the authors' name to assign a unique ID for every author, to account for different naming styles used by authors over time. For naming disambiguation, we use edit distance between strings to cluster similar names and then check for neighbourhood overlap in the co-authorship network. Names with small edit distance and high neighbourhood overlap were grouped together and manually checked for uniqueness using information from online databases. This process resulted in 8084 unique Indian authors.

Then, we constructed bipartite graphs $B = (U, V, E)$ for every year, from 1970 to 2013, where V is the set of papers, U is the set of authors and E is the set of edges connecting nodes from U and V. Each graph $B(t)$ at time t is cumulative, storing all the information until time t. From each $B(t)$ we projected the bipartite network to get a weighted, cumulative, undirected co-authorship network. We also constructed cumulative directed citation networks for every year using the paper IDs of Indian publications from the APS citations dataset. We illustrate the process of creating these networks in Fig. 1. The ordering of node labels in the projected co-authorship networks was kept consistent for all calculations.

Using the above graphs, we constructed our data matrices for co-authorship and citation networks as follows.

Data matrices

In order to aid our analysis, we created two types of matrices, one for the co-authorship networks and the other for the citation networks, for each of the 44 years. The matrices have size $N \times N$, where $N = 8084$ is the number of unique authors in the whole dataset.

The elements of the first matrix type, $D(t)$, are given by

$$d_{ij}(t) = \begin{cases} d \text{ if there is a path of length } d \text{ from } i \text{ to } j \\ 0 \text{ otherwise,} \end{cases}$$

such that the matrix $D(t)$ captures the distances between all possible $\binom{N}{2}$ pairs of authors in the network at time t. For $d_{ij} = 0$, it can mean that the nodes do not exist in the network at that time or that they are not connected via any path.

The second matrix type, $C(t)$, stores the citations exchanged between papers written by i and j *before* a given time t. That is, $c_{i \leftarrow j}(t)$ is the cumulative number of times that j cites i until that particular year.

To see the ageing effect in collaboration between authors, we trace the history of co-authorship events. First, we extract the distance $d_{ij} = 1$ collaborations at the end of the study period (i.e. the edges in the 2013 co-authorship graph). Next, we trace the presence of an edge between i and j (contributing authors) in reverse order. The time when the edge first appears is marked T_c, which is the year of first collaboration. Then, we check for the presence of i and j in the network before T_c until we find T_0, the first year in which authors i and j both are present in the network. For each year from T_0 until 2013 we record the number of citations exchanged by that pair of authors, before and after their first collaboration. A diagram illustrates this method of tracing the history of collaboration between a pair of nodes in Fig. 2.

In order to compare the history of citation and co-authorship for all pairs of authors, we shift the time series on the x-axis for every pair such that T_c, the year of the first collaboration between two authors, is centred on zero. We then remove the pairs that
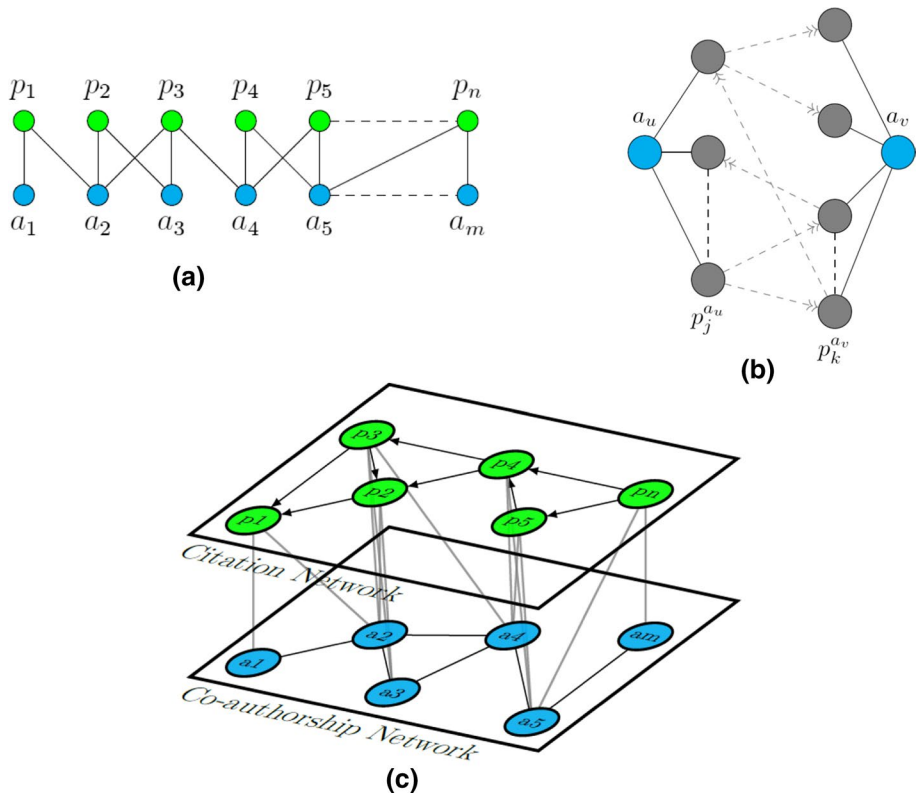
Fig. 1 Diagrams of interdependent citation and co-authorship networks. **a** Bipartite network between papers p_i and authors a_i constructed cumulatively. **b** The exchange of citations (dotted arrows) between papers $p_j^{a_u}$ by author a_u and $p_k^{a_v}$ by author a_v. **c** A multilayer representation of the interdependence between the citation network and the projected co-authorship network constructed from (**a**)

have $T_0 = T_c$. We do this to remove authors who first appear in the network with a shared publication. We exclude these authors since they will not have any citing history before their first collaboration. The remaining are pairs of authors who took at least 1 year to collaborate after both were present in the network.

In addition to the history of citations, we calculated the fraction of self-citations over the total citations exchanged between author pairs. For every pair of authors that co-authored a paper, we count the number of outgoing (given) and incoming (received) self-citations. These were calculated as follows:

While calculating the total citations exchanged between authors i and j we count a citation as an outgoing (given) self-citation if, i cited a paper by j and was also present in the contributing authors of that paper. Similarly, if i received a citation from a paper by j and was also a contributing author in the paper, we counted the citation as an incoming (received) self-citation.

Fig. 2 Schematic diagram of the method of counting the number of citations exchanged between pairs of authors before and after their first collaboration. Here we are interested to see the difference in mutual citations before and after T_C. Author pairs that have $T_0 = T_C$ are ignored here
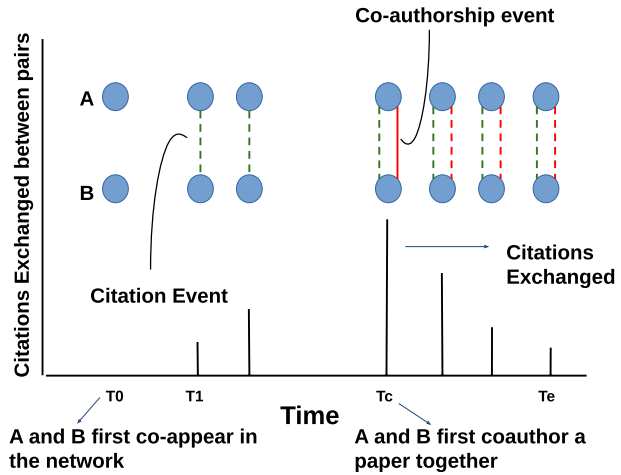
Calculations

The data matrices $D(t)$ and $C(t)$ store the information of the distance and the citations exchanged between all possible pairs of nodes for the co-authorship and citation networks, respectively, for every time step (which is a year in our case). These matrices enable us to calculate any changes in distance or citations exchanged from 1 year to another. Then, to address the research questions mentioned in the Introduction, we define our calculations based on the different situations that each possible pair of nodes can exhibit in the networks. More specifically, we calculate the following, at every time step t:

1. What fraction of authors exchange citations but do not co-author?

   (a) We count pairs that exchange citations ($c_{ij} + c_{ji} \neq 0$) and have a connected path in the co-authorship network ($d_{ij} > 1$).
   (b) We count pairs that exchange citations ($c_{ij} + c_{ji} \neq 0$) but are not connected in the co-authorship network $d_{ij} = 0$. These are the pairs that are aware of each other's work via citations but do not have any direct or indirect connection with each other via co-authorship.

2. How are citations exchanged between co-authors?

   (a) We count pairs that co-author ($d_{ij} = 1$) but do not exchange citations ($c_{ij} + c_{ji} = 0$).
   (b) We count pairs who are co-authors ($d_{ij} = 1$) and exchange citations ($c_{ij} + c_{ji} \neq 0$).

3. How do the statistics in 1 vary with network distances between authors? We count pairs that exchange citations ($c_{ij} + c_{ji} \neq 0$) for different distances d_{ij}.
4. How does receiving a new citation affect the likelihood of an author creating a new link in the co-authorship or the citation network?
   (a) Response by authors, in terms of citations (how they cite back), to other authors who cited them, and to the total citations they received. For every author i in the

citation matrix $C(t)$, we define N_i as the number of authors that cite i. Among N_i, n_i is the number of authors whom i cites back. Similarly, $C_{in}^{N_i}$ are the total citations received by i from the set N_i of authors and $C_{out}^{n_i}$ are the total number of citation given out by i to the set of n_i authors. The response by an author is calculated as: (a) n_i/N_i—the response to citing authors; and b) $C_{out}^{n_i}/C_{in}^{N_i}$—the response to citations received.

5. What is the relationship between the probability of citations and network distance?

(a) Correlation between citations exchanged and network distance. In order to define the probability of citations between authors, we take $\mathcal{P}_{ij}^C(t)$ as the probability that j cites i at time t, given by

$$\mathcal{P}_{ij}^C(t) = f_c^i(t)c^j(t^*), \tag{1}$$

where $f_c^i(t)$ is the fraction of citations i has received prior to t, and $c^j(t^*)$ is the fraction of citations j gives out in the year between $t-1$ and t. We also define the mean probability

$$\mathcal{P}^C(t) = \langle \mathcal{P}_{ij}^C(t) \rangle. \tag{2}$$

We calculate $f_c^i(t)$ and $c^j(t^*)$ from our data matrices according to

$$f_i^c(t) = \frac{\sum_j c_{ij}(t)}{\sum_i \sum_j c_{ij}(t)}, \tag{3}$$

$$c^j(t^*) = \frac{\sum_i c_{ij}(t^*)}{\sum_i \sum_j c_{ij}(t^*)}, \tag{4}$$

where

$$c_{ij}(t^*) = c_{ij}(t) - c_{ij}(t-1). \tag{5}$$

Our reason behind this approach is twofold: (1) Popular authors (or papers) have a greater tendency to get cited ($f_c^i(t)$)—the frequently observed preferential attachment phenomenon; and (2) if an author (paper) is giving out more citations it uniformly increases the probability of other authors (papers) getting cited ($c^j(t^*)$). It should be noted that this definition is independent of the relationship between authors in the co-authorship network. This approach will serve as the null model for our subsequent comparisons since it calculates probability distributions for citations without accounting for co-authorship distance.

Equation (1) estimates the probability of a directed edge between authors in the citation graph, given no information other than the number of citations exchanged between authors recorded in $C(t)$. Next, we need to define the relations to compare the empirical observations from the citation and co-authorship networks with the null model. We use Bayes' theorem to construct probability relationships. This helps us to put constraints in our observations that will help us to highlight the dependency of citations on the shortest path between authors in the co-authorship network.

If T_0 is the time of first co-appearance of i and j and $P(d=1)$ is the probability of them co-authoring, we have

$$P(d = 1|C^{>0}(t)) = \frac{P(C^{>0}(t)|d = 1) \times P(d = 1, t)}{P(C^{>0}(t))}, \tag{6}$$

where

$$P(C^{>0}(t)) = \sum_{k=0}^{\infty} P(C^{>0}(t)|d = k) \times P(d = k) \tag{7}$$

Equation (6) is the empirical probability of observing pairs of authors connected by a path of length one, given that they exchange non-zero citations ($C^{>0}(t)$), normalized by the probability of non-zero citations for pairs at all possible shortest path lengths (Eq. (7)). Therefore, for any distance $d = k$, we have

$$P(d = k|C^{>0}(t)) = \frac{P(C^{>0}(t)|d = k) \cdot P(d = k)}{P(C^{>0}(t))} \tag{8}$$

If we reverse the relationship using Bayes' rule, the empirical probability of observing pairs with non-zero citations between them, given they are at a distance $d = k$ is calculated according to

$$P(C^{>0}(t)|d = k)$$
$$= \frac{P(d = k|C^{>0}(t)) \cdot P(C^{>0}(t))}{P(d = k|C^{>0}(t))P(C^{>0}(t)) + P(d = k|C^0(t))P(C^0(t))} \tag{9}$$

The denominator in Eq. (9) normalizes over pairs that either exchange citations ($C^{>0}(t)$) cite or do not cite each other ($C^0(t)$) given a network distance $d = k$.

6. What is the waiting time distribution for consecutive co-authorship events and consecutive co-citation events?

   (a) *Co-authorship events* The co-authorship networks are weighted, undirected networks constructed cumulatively. Therefore, every time an author publishes a paper with a co-author, the weight of the edge between them in the co-authorship network changes. We record the time Δt it takes for this change to happen. We do so for all pairs of co-authors in the network over time.
   (b) *Co-citation events* For every pair ij in the citation matrix $C(t)$ we record the time Δt it takes for a change in the value of $(c_{ij} + c_{ji})$ over the whole time period.

After defining our data matrices, $C(t)$ and $D(t)$, and the distinct relations between citations exchanged and distance in the co-authorship networks, we turn our focus to the understanding of the interdependence between the co-authorship and the citation networks. That is what we address in the next section.

Results and discussions

Using the data matrices $C(t)$ and $D(t)$, described above, we count the number of pairs for different citation and co-authorship distance relations as the networks evolve with time. To understand the interdependence between the citation network and its associated co-authorship network we calculate the citations exchanged between pairs by splitting them into three groups:

1. $D_{ij} > 1$: Pairs that are connected with a shortest path length $d > 1$ in the co-authorship network (Fig. 3a),
2. $D_{ij} = 1$: Pairs that co-author a paper together (Fig. 3b), and
3. $D_{ij} = 0$: Pairs that are not connected in the co-authorship network (Fig. 3c).

For each group, we separately count the number of pairs of authors that do not exchange any citation (orange line in Fig. 3) and pairs that have non-zero citations shared between themselves (blue line in Fig. 3). For all cases, as the citation and co-authorship networks grow, the fraction of pairs that do not have any citations between them is larger than the fraction of pairs that do exchange citations.

The contributions of each of the three groups defined, relative to the total number of citations are shown in Fig. 4. First, the light green region in Fig. 4 is the fraction of citations exchanged between pairs that have distance $d > 1$ in the co-authorship networks. Even though the number of such pairs is a small fraction in the co-authorship networks (blue line in Fig. 3a), they still contribute significantly to the total number of citations. Second, pairs that co-author are responsible for most of the citations (blue region in Fig. 4), which reflects the importance of an authors' collaborators to the number of citations received. Third, disconnected pairs exchange a tiny fraction of total citations between them (sky blue region in Fig. 4), showing a decreasing trend, until it almost vanishes in the last years.

The behaviour described above could be a consequence of our choice of the dataset. Since we focus on a small fraction of the total number of publications (those from Indian authors) in the global APS network, authors are expected to be well connected and closely
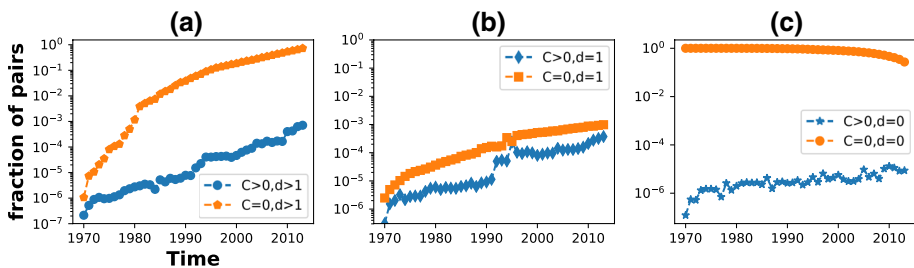
Fig. 3 Number of pairs of authors with different citation and distance relations changing with time. Each sub-figure is for pairs at different shortest path length in the co-authorship networks. **a** Pairs that are connected with path length greater than one, **b** pairs that co-author a paper, and **c** pairs that are not connected at all. The orange line represents pairs that do not exchange any citations while the blue line is for pairs that exchange at least one citation. The number of pairs at each time is divided by total possible pairs. I.e. $\binom{N}{2}$ where $N = 8084$. The blue line in each figure suggests that a tiny fraction of the total possible pairs in the co-authorship network are responsible for all the citations observed. We are interested in seeing the patterns in citations between these pairs. (Color figure online)
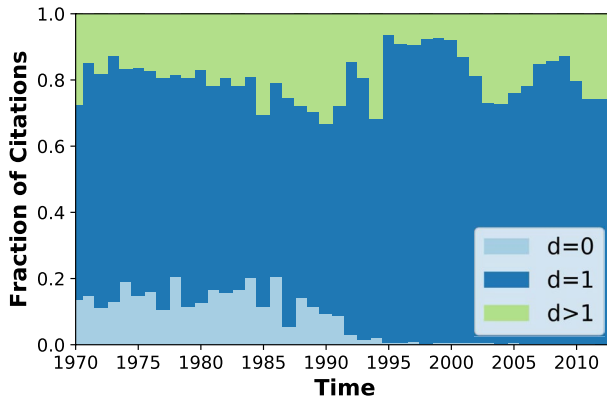
Fig. 4 The fraction of citations exchanged between pairs at different network distances over time distributed between pairs at different path lengths as in (a), (b) and (c) in Fig.3. The figure suggests that the majority of citations are exchanged between co-authors (blue region) $d = 1$. Distant connections $d > 1$ (green region) also contribute to the total citations however disconnected pairs $d = 0$ (sky blue region) contribute in the beginning but decrease to an almost negligible effect in later times. (Color figure online)

citing each other. As most of the pairs are connected by a path in a growing co-authorship network—represented by the orange line in Fig. 3a approaching one—and are aware of each other in the network (the decrease in the orange line in Fig. 3c) the citations by distant pairs decreases. The trend in the number of co-authoring pairs that exchange citations shows an interesting sudden jump in the mid-1990s. We believe this trend is due to the increasing number of nodes in the co-authorship network. In the beginning, there were very few nodes (authors) in the network, most of which appeared as co-authoring pairs, which explains the initial increase. Post-1993 (the blue line in Fig. 3b) we notice a sudden increase in the number of such pairs. This sudden change is because of the introduction of papers with a high number of authors in that period. These papers lead to large cliques (totally connected sub-graphs) in the co-authorship network. Most of these papers are published by large collaboration groups often having multiple common authors in their publications and with many Indian authors being part of such groups. For example, the papers Abachi et al. (1995) and Abelev et al. (2010) have 395, and 383 authors respectively. Even one citation shared between such papers would dramatically inflate the number of citations exchanged, due to the large size of the induced co-author cliques. We feel that these fluctuations observed in citation pattern should be unique to Physics. This is because other fields do not observe such an abundance of papers with very high numbers of authors (Vasques Filho and O'Neale 2020).

In the above calculations, we counted the total citations exchanged between the pairs in the co-authorship network for different network distances, normalized by the total number of possible pairs in the co-authorship network. The citation count included both incoming and outgoing citations. To measure the response of authors to incoming citations, we split our calculations into two parts. First, we calculate the average fraction of outgoing citations from authors for every citation received by them. We observe that over time people tend to cite more and more articles in their work; hence, we see an increasing trend in response to citations (the orange curve in Fig. 5). Second, we calculate the average fraction of incoming citations that an author responds to by subsequently citing the author who initially cited her (the blue curve in Fig. 5). When the co-authorship network is in its initial phase, with

Fig. 5 The increasing mean probability for authors to give back a citation for every citation they receive (orange squares) indicates the tendency of authors to cite more, i.e. longer reference list. On the other hand, the decreasing mean probability for authors to cite back someone who cited them in the past shows the aging effect in mutual citations. (Color figure online)
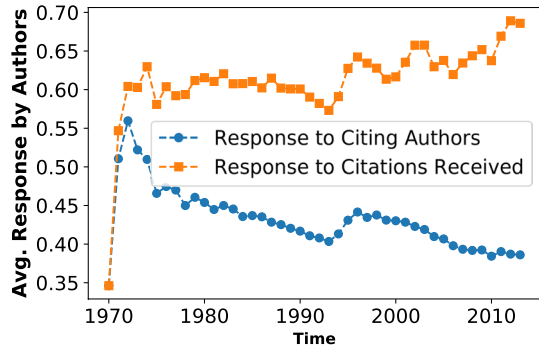

a small number of researchers, most co-author pairs cite each other. Hence, we observe high citation reciprocity at initial times. As the network grows, the distribution of citations becomes more heterogeneous as some authors receive more citations than others (authors of influential papers receive many citations). Besides, authors that are no longer publishing cannot reciprocate anymore but still receive citations. Thus, more citations are given out than received on average, which results in a decreasing trend in citation reciprocity (the blue line in Fig. 5). The sudden increase in reciprocity in the mid-1990s is due to the citations exchanged between papers with a large number of authors, which, as aforementioned, started to appear around that time.

So far, our observations give a macroscopic understanding of the interdependence of simultaneously growing citation and co-authorship networks. To probe that further, we make more elaborate calculations to see the effect of co-authorship network distance on the citations exchanged. In the interest of this objective, we ignore pairs that do not exchange any citations (orange line in Fig. 3) from our subsequent analysis.

The number of citations exchanged by pairs of authors ij at a distance d_{ij} in the co-authorship network decreases rapidly with increasing co-authorship distance (Fig. 6). We plot this relation for networks at different times (1990, 2000, and 2013) to show that the trend is consistent as the network evolves. The average citations between pairs (Fig. 7a) displays a significant difference in the temporal trends for different network distances. For direct collaborations (pairs with $d_{ij} = 1$), the rate of citation exchange increases with time, more rapidly after 1995. This increase is likely due to a sudden

Fig. 6 Average total number of citations between pairs of authors versus distance d between pairs in the co-authorship network (for years till 1990, 2000 and 2013). Consistency in the trend for every time period indicates an interdependence between mutual citations between pairs and their co-authorship network distance. Average citations fall rapidly up to $d \leq 3$ and have a similar trend for longer distances
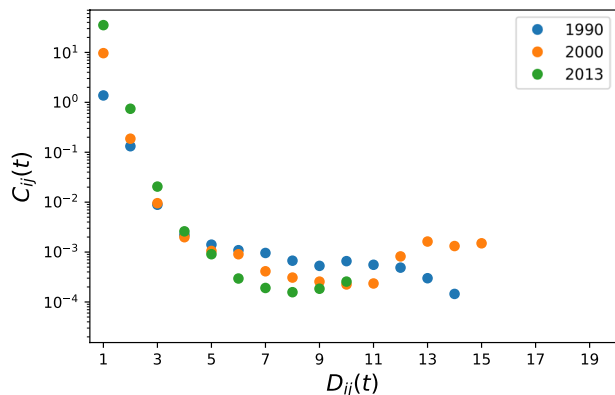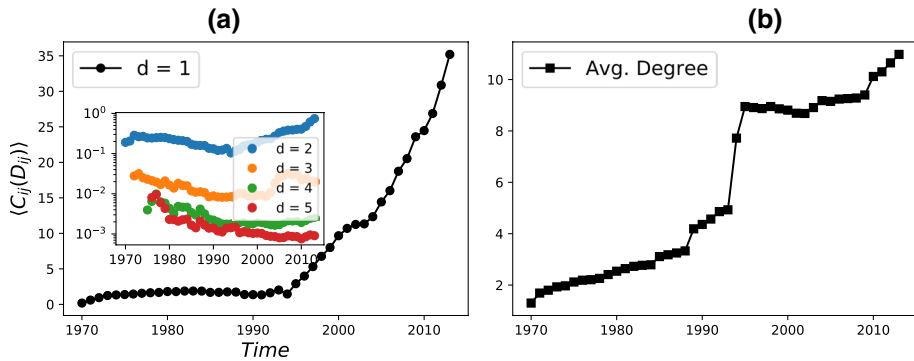
Fig. 7 **a** Change in the average number of citations exchanged between pair i, j at distance $d = 1$ with time. The sudden increase in citations for pairs at $d = 1$ post 1994 can be attributed to sudden increase in the degree in (**b**) i.e. number of collaborators. The inset in (**a**) shows the variation of citations for higher network distances. We see a significant difference (almost by an order of magnitude) between $d = 1, 2$ and 3. For $d > 3$ the trend is similar. Here we have plotted only up to $d = 5$ for clarity

increase in the number of collaborators (as seen by the change in the average degree of nodes in Fig. 7b). The average number of citations differ roughly by one order of magnitude for co-authorship distances of $d_{ij} = 1, 2$ and 3; for larger distances ($d \geq 4$), the average number of citations exchanged is very low, with similar trends, as shown in Figs. 7a and 7a (inset).

The change in the strength of collaboration between two authors i and j is measured by the total citations exchanged between i and j (scatter plot Fig. 8a) before and after the period T_c of first collaboration (black line in Fig. 8a), including cases where i and j cite their previous papers. Co-author pairs exhibit an interesting citing pattern—the number of citations shows a steep rise after the first co-authorship event and then decays with time, indicating an ageing effect.

Interestingly, the peak for this distribution is within 5 years of T_c. When averaged over all times, we notice that the decaying trend is well fitted by a Weibull distribution, $f(t) = \frac{b}{n}(\frac{t}{n})^{b-1}e^{-(\frac{t}{n})^b}$, as was noted in Börner et al. (2004).
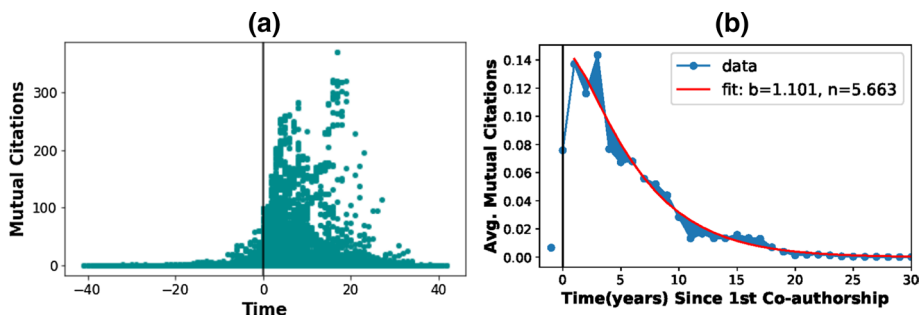

Fig. 8 **a** Scatter plot of citations exchanged before and after time of first co-authorship event (black vertical line) between authors. Authors exchange more citations immediately after they co-author a paper together. This mutual citation then decreases with time. **b** The average number of citations at each time (measured relative to the time of initial co-authorship) fitted to a Weibull distribution. The trend indicates an aging effect in mutual citation between pairs of authors

Fig. 9 Trends in the fraction of self-citations given (orange) and received (blue) over total outgoing and in-coming citations, respectively. The trend lines in both cases indicate pairs of authors exchanging a high portion of self-citations. (Color figure online)
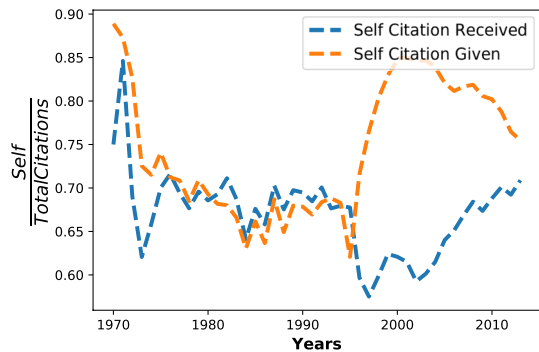

Fig. 10 Average number of authors and references per paper over time from our data-set. An increasing trend in both cases indicates a correlation. The average number of references is relatively less as this only includes references within American Physical Society (APS) journals. (Color figure online)
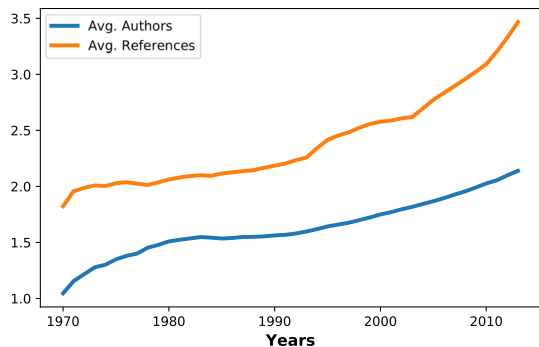

Another interesting observation was noted when we compared the fraction of self-citations over total citations exchanged between pairs of authors who collaborate. In Fig. 9 we see a sudden change in the trends of total self-citations given and received in the mid-'90s. This change can be explained by the appearance of papers with a large number of authors, arising from large collaborative research projects. Such projects are likely to produce papers that cite other publications related to the same project and to have a high number of overlapping authors. This leads to the observed large number of self-citations. While all authors do cite others in their papers, few of them receive citations. Therefore, while the outgoing self-citations increase rapidly, the incoming self-citations observe a dip. However, over time as authors start to receive citations from their close collaborators (co-authors) we observe an increasing trend in the self-citations received. In total 71% of the total citations received and 75% of the citations given out are self-citations.

It has also been observed that over time the number of references in a paper have increased (orange line in Fig. 10). There are a number of potential explanations for this trend. One is that, there is an increase in the heterogeneity of references as a result of increasing research complexity and interdisciplinarity. Another possible explanation is that improvements in bibliographic technology resulting from online research resources (Web of Science, Scopus, Google Scholar, and the like) and increasing online journal access have made it easier for researchers to find and cite large numbers of articles. A third potential explanation is that the increasing number of references is a function of the increase in the number of publications—authors cite more articles simply because there are more articles to be cited. The reality is likely to be some combination of these, and other, effects.
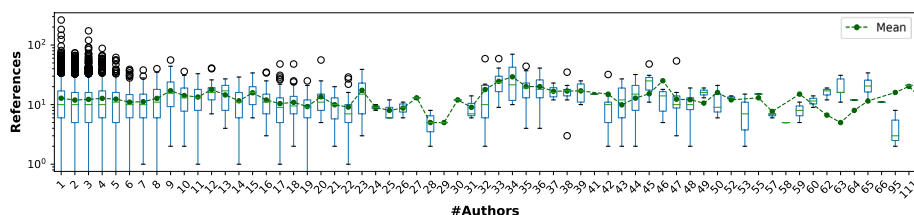
Fig. 11 Distribution of number of references in papers with n authors. The flatness in the trend of the average indicates that even if the average references and average authors in papers have a correlated trend, they do not have any causal relationship

Fig. 12 Waiting time distributions between consecutive citations and consecutive co-authorship events for pairs of authors. the black dashed line indicates the 5-year mark, before which more than 95% of citation and co-authorship events occur
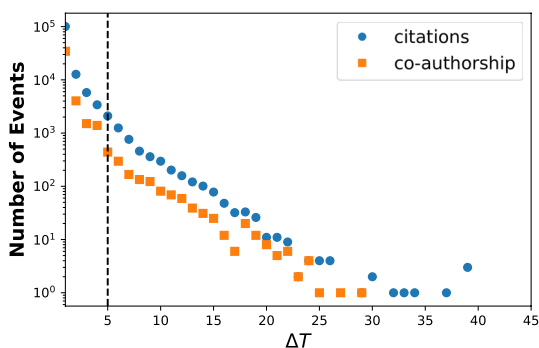

We also investigated whether this effect could be a result of the increasing mean number of authors per paper, that is, the hypothesis that each author suggests a similar number of references to cite (blue line in Fig. 10). The plot in Fig. 10 indicates a positive correlation between the two. However, we argue that this is not an indicator of causality for two reasons. First, the total references in our data-set are only the ones that are within American Physical Society (APS) journals. This excludes a good portion of the total references in a paper, hence results are indicative but not conclusive. Second, when we looked at the distribution of the number of references in n-authored papers where n is the number of authors in the paper as shown in Fig. 11, we observed a consistent mean of the distribution over n. This indicates no strong relation between the increasing number of authors with the increased number of references per paper.

Next, we calculate the waiting time (ΔT) distribution for consecutive citations between pairs of authors and consecutive co-authorship events. Both follow a similar trend (Fig. 12) with the majority of co-authorship and citation events (95%) happening within the first 5 years of the initial event (black dotted line in Fig. 12).

Finally, we calculate the empirical probabilities of authors acquiring citations, firstly with the null model (Eq. (1)) and subsequently with the empirical probabilities derived using Bayes' formalism (Eqs. (6–9)). The latter accounts for co-authorship distance in determining the probability of citations exchanged between pairs (Fig. 13). We notice that the null model, which is proportional to the popularity of the author (paper), and the number of citations given out by the citing author (paper) are not sufficient to explain the observed behaviour of citation exchange. The citing patterns significantly differ for different network distances between author pairs. The probability of citations between pairs at $d_{ij} = 1$ (blue line in Fig. 13) closely follows the null model, while pairs

Fig. 13 Comparing the probability obtained from Bayes' formalism versus the hypothesis for different network distances. A clear difference in probabilities for different distances between pairs indicates a distance dependence in citation between authors. (Color figure online)
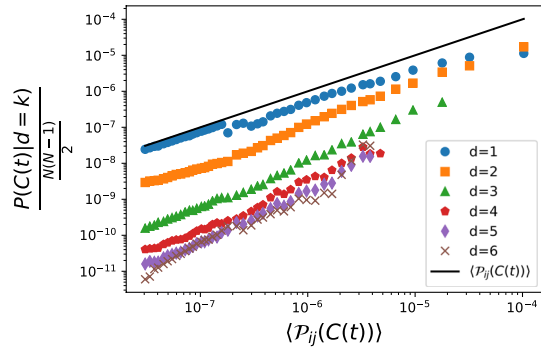
at greater distances significantly differ from it. From this, we infer that most of the overall citation behaviour is explained by only considering those citations that come directly from co-authors with $d_{ij} = 1$. This conclusion is also made evident by the blue region in Fig. 4 where citations from co-authors contribute to most of the total number of citations. Distinct probabilities of citations at different co-authorship distances (Fig. 13) and the decay in average citations with higher network distances (Fig. 6) indicate interdependence between citing patterns and co-authorship network distances, confirming our hypothesis.

By splitting our analysis into different research questions, we were able to explore both macroscopic and microscopic trends in citing patterns between authors as they appear in the associated co-authorship network.

In our macroscopic approach, we first count connected pairs, co-authors and disconnected pairs in Fig. 3 and their contributions to total citations exchanged in Fig. 4 to find that

- A tiny fraction of possible pairs are connected with distance $d_{ij} > 1$ in the co-authorship network but still have a significant contribution to the overall number of citations.
- Co-authors are a tiny fraction of the total possible number of pairs but account for most of the citation exchanges observed.
- Disconnected pairs contribute to citations at the beginning of the network, but their contribution becomes negligible as the network grows and becomes more connected.

Besides interactions between pairs, we also investigated the average probability of an author citing back an author that has cited her (the blue line in Fig. 5) and the average probability of an author giving back a citation for every citation received (the orange line in Fig. 5). The trends indicate that while, over time, the ratio of outgoing to incoming citations per author has increased. However, the ratio of pairwise outgoing citations that reciprocate citing authors decreases over time. In Fig. 8, the plot of citations exchanged between authors exhibits a sudden increase when they co-author a paper and then decays, which is consistent with the ageing effect in citations and collaboration reported by earlier studies. A similar effect is observed in Fig. 12 where 95% of consecutive co-authorship and citation events happen within the first 5 years of a first co-authorship event.

On the other hand, in microscopic calculations, we first calculate the average citations shared between author pairs at all possible network distances for networks at

different points in time (Fig. 6) and for all time steps (Fig. 7—plotted only up to $d = 5$ for clarity). The number of average citations shows a steep decay up to $d \leq 3$ and then is almost stable for longer distances. That is, pairs that are more than distance three apart in the co-authorship network have a similar (and minimal) effect on citation patterns. To confirm the interdependence between the citation network and the associated co-authorship network, we formulate a null model for the probability of author j citing author i in Eq. (1) and then using Bayes' formalism (Eqs. (6–9)), we explicitly show that the null model is indeed insufficient to explain the citing patterns. There is a significant effect caused by the distance between pairs of authors in the co-authorship network over their citing patterns. The effect is most dominant for immediate co-authors (Fig. 13).

Conclusion

The main contribution of this study lies in a rigorous and comprehensive analysis that probes the relations between the distances in the co-authorship network and the citation patterns in the citation network. We do this for all possible pairs of researchers from the citation and co-authorship networks constructed. For all pairs of authors, we observe the number of citations exchanged between them as a function of distance in the co-authorship network, as it evolves. We find that co-authors dominate the citation patterns in our networks with a heavy bias ($\approx 75\%$) towards self-citations. The remainder of the citations were mostly between pairs that have a short ($d \leq 3$) connected path in the co-authorship network; the average number of citations exchanged decays with increasing co-authorship distance. Pairs with distances $d > 3$ have a relatively small contribution to the citations. Disconnected pairs of authors make a small contribution to the citations in the initial years of the network but quickly become almost negligible, as the network grows to be more connected over time. We also highlight the underlying ageing effect in mutual citations and collaborations.

Most of the citations happen within three degrees of separation in the co-authorship network of our dataset. This pattern indicates that authors mostly cite their co-authors and the collaborators of their co-authors. Since we study researchers affiliated to Indian institutes, these citations exchanged would be between authors working in similar research topics with possible connections in the co-authorship network. In short, the similarity in the research topic and affinity towards close collaborators would be the major effects driving the citation patterns in our case. However, as some authors (or papers) gain more citations over time, the dynamics for the top-cited authors (papers) are likely to differ from those of the majority. In such cases, distant or disconnected authors will also have a significant contribution to the total number of citations of the author (or paper). Therefore, to measure the impact of a paper concerning citations received, we should have measures that account for this possible bifurcation in citing patterns.

The main advantage we took from our dataset was in calculating pairwise interactions. Even so, we realize that our dataset is not entirely comprehensive, as it considers only Indian authors and APS publications, hence our results might show small variations when calculated for larger datasets. However, we believe that this difference should not be significantly large, assuming that pairwise interactions between authors would be similar irrespective of the size of the network.

We reflect on the opinion that most real-world networks can be viewed as interdependent multi-layer networks, with networks of scientific collaborations as an example of

this. This interdependence is critical when studying the dynamics of such networks. Our analysis explicitly shows that connected paths in one network (co-authorship) impact the structure of the other (citation). The dominance of pairs close in distance highlight the importance of an authors' neighbourhood in her citing patterns which, in turn, can be used to explain the patterns in the flow of ideas and information in the scientific ecosystem. Results from this study can be used in the development of more sophisticated models to investigate the spread of scientific knowledge. We believe that to understand the true nature of research collaboration, it is essential to consider both co-authorship and citation networks simultaneously.

References

Abachi, S., Abbott, B., Abolins, M., Acharya, B. S., Adam, I., Adams, D. L., et al. (1995). Top quark search with the d0 1992–1993 data sample. *Physical Review D*, *52*(9), 4877.

Abelev, B. I., Aggarwal, M. M., Ahammed, Z., Alakhverdyants, A. V., Anderson, B. D., Arkhipkin, D., et al. (2010). Identified particle production, azimuthal anisotropy, and interferometry measurements in au+ au collisions at s nn= 9. 2 gev. *Physical Review C*, *81*(2), 024911.

Amblard, F., Casteigts, A., Flocchini, P., Quattrociocchi, W., & Santoro, N. (2011). On the temporal analysis of scientific network evolution. In *2011 international conference on computational aspects of social networks (CASoN). IEEE* (pp. 169–174).

Barabási, A.-L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A*, *311*(3–4), 590–614.

Biscaro, C., & Giupponi, C. (2014). Co-authorship and bibliographic coupling network effects on citations. *PLoS ONE*, *9*(6), e99502.

Börner, K., Maru, J. T., & Goldstone, R. L. (2004). The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(suppl 1), 5266–5273.

Chen, Y., Börner, K., & Fang, S. (2013). Evolving collaboration networks in scientometrics in 1978–2010: A micro–macro analysis. *Scientometrics*, *95*(3), 1051–1070.

Ding, Y. (2011). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of Informetrics*, *5*(1), 187–203.

Dong, Y., Ma, H., Shen, Z., & Wang, K. (2017). A century of science: Globalization of scientific collaborations, citations, and innovations. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1437–1446). ACM.

Fiala, D., Tutoky, G., Koncz, P., & Paralič, J. (2015). Ageing of edges in collaboration networks and its effect on author rankings. *Acta Polytechnica Hungarica*, *12*(6), 149–160.

Glänzel, W., & Thijs, B. (2004). Does co-authorship inflate the share of self-citations? *Scientometrics*, *61*(3), 395–404.

González-Teruel, A., González-Alcaide, G., Barrios, M., & Abad-García, M. (2015). Mapping recent information behavior research: An analysis of co-authorship and co-citation networks. *Scientometrics*, *103*(2), 687–705.

Gualdi, S., Yeung, C. H., & Zhang, Y.-C. (2011). Tracing the evolution of physics on the backbone of citation networks. *Physical Review E*, *84*(4), 046104.

Guo, L., Cai, X., Hao, F., Mu, D., Fang, C., & Yang, L. (2017). Exploiting fine-grained co-authorship for personalized citation recommendation. *IEEE Access*, *5*, 12714–12725.

Hajra, K. B., & Sen, P. (2005). Aging in citation networks. *Physica A*, *346*(1–2), 44–48.

Higham, K. W., Governale, M., Jaffe, A. B., & Zülicke, U. (2017). Unraveling the dynamics of growth, aging and inflation for citations to scientific articles from specific research fields. *Journal of Informetrics*, *11*(4), 1190–1200.

Huang, J., Zhuang, Z., Li, J., & Giles, C. L. (2008). Collaboration over time: Characterizing and modeling network evolution. In *Proceedings of the 2008 international conference on web search and data mining* (pp. 107–116). ACM.

Kas, M., Carley, K. M., & Carley, L. R. (2012). Trends in science networks: Understanding structures and statistics of scientific networks. *SNAM*, *2*(2), 169–187.

Keegan, B., Horn, D., Finholt, T. A., & Kaye, J. (2013). Structure and dynamics of coauthorship, citation, and impact within CSCW. arXiv preprint arXiv:1307.7172.

Lande, D. V., & Andrushchenko, V. B. (2016). Formation of subject area and the co-authors network by sounding of Google Scholar citations service. arXiv preprint arXiv:1605.02215.

Martin, T., Ball, B., Karrer, B., & Newman, M. E. J. (2013). Coauthorship and citation patterns in the physical review. *Physical Review E*, *88*(1), 012814.

Newman, M. E . J. (2001a). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, *64*(1), 016132.

Newman, M. E. J. (2001b). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(2), 404–409.

Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(suppl 1), 5200–5205.

Pan, R. K., Petersen, A. M., Pammolli, F., & Fortunato, S. (2018). The memory of science: Inflation, myopia, and the knowledge network. *Journal of Informetrics*, *12*(3), 656–678.

Sarigöl, E., Pfitzner, R., Scholtes, I., Garas, A., & Schweitzer, F. (2014). Predicting scientific success based on coauthorship networks. *EPJ Data Science*, *3*(1), 9.

Shi, X., Tseng, B., & Adamic, L. (2009). Information diffusion in computer science citation networks. In *Third international AAAI conference on weblogs and social media*.

Sin, S.-C. J. (2011). International coauthorship and citation impact: A bibliometric study of six lis journals, 1980–2008. *Journal of the American Society for Information Science and Technology*, *62*(9), 1770–1783.

Sinatra, R., Wang, D., Deville, P., Song, C., & Barabási, A.-L. (2016). Quantifying the evolution of individual scientific impact. *Science*, *354*(6312), aaf5239.

Singh, C. K., & Jolad, S. (2019). Structure and evolution of indian physics co-authorship networks. *Scientometrics*, *118*(2), 385–406.

Tol, R. S. J. (2011). Credit where credit's due: Accounting for co-authorship in citation counts. *Scientometrics*, *89*(1), 291.

Tomassini, M., & Luthi, L. (2007). Empirical analysis of the evolution of a scientific collaboration network. *Physica A*, *385*(2), 750–764.

Tutoky, G., & Paralič, J. (2011). Time based modeling of collaboration social networks. In *International conference on computational collective intelligence* (pp. 409–418). Springer.

Vasques Filho, D. (2018). *Structure and dynamics of social bipartite and projected networks*. PhD thesis, The University of Auckland.

Vasques Filho, D., & O'Neale, D. R. (2020). Structure dynamics of evolving scientific networks. arXiv, pages arXiv-2005.

Wallace, M. L., Larivière, V., & Gingras, Y. (2012). A small world of citations? The influence of collaboration networks on citation practices. *PLoS ONE*, *7*(3), e33339.

Xu, G., He, X., & Butler, A. (2017). What comes first the co-authorship network or the citation?

Yan, E., & Ding, Y. (2012). Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. *Journal of the American Society for Information Science and Technology*, *63*(7), 1313–1326.

Zhan, C., & Tse, C. K. (2017). A network model for growth of publications and citations. *Journal of Complex Networks*, *5*(2), 303–314.

Affiliations

Chakresh Kumar Singh[1] [ID] **· Demival Vasques Filho[2,3] · Shivakumar Jolad[4] ·
Dion R. J. O'Neale[2]**

Demival Vasques Filho
vasquesfilho@ieg-mainz.de

Shivakumar Jolad
shivakumar.jolad@flame.edu.in

Dion R. J. O'Neale
d.oneale@auckland.ac.nz

[1]   Indian Institute of Technology, Gandhinagar 382355, India

[2]   Te Pūnaha Matatini, Department of Physics, University of Auckland, Private Bag 92019,
       Auckland, New Zealand

[3]   Leibniz-Institut für Europäische Geschichte, Alte Universitätsstraße 19, 55116 Mainz, Germany

[4]   FLAME University, Pune, India