



RESEARCH ARTICLE

# Measuring research interest similarity with transition probabilities

Attila Varga<sup>1</sup> , Sadamori Kojaku<sup>2</sup> , and Filipi Nascimento Silva<sup>3</sup> 

<sup>1</sup>Corvinus Institute for Advanced Studies and Institute of Social and Political Sciences, Corvinus University of Budapest, Budapest, Hungary

<sup>2</sup>Thomas J. Watson School of Engineering and Applied Science, Binghamton University, Binghamton, NY, USA

<sup>3</sup>Indiana University Network Science Institute, Indiana University Bloomington, Bloomington, IN, USA

an open access  journal



**Keywords:** citation networks, paper similarity, research problem choice, transition probability

## ABSTRACT

We introduce a family of paper and author similarity measures based on the concept that papers are more similar if they are more likely to be retrieved during a literature search following backward and forward citations. As this browsing process resembles a walk in a citation network, we operationalize the concept using the transition probability (TP) of random walkers. The proposed measures are continuous and symmetric, and can be implemented on any citation network. We conduct validation tests of the TP concept and other extant alternatives to gauge which metric can classify papers and predict future coauthors most consistently across different scales of analysis (coauthorships, journals, and disciplines). Our results show that the proposed basic TP measure outperforms alternative metrics such as personalized PageRank and the node2vec machine-learning technique in classification tasks at various scales. Additionally, we discuss how publication-level data can be leveraged to approximate the research interest similarity of individual scientists. This paper is accompanied by a Python package that implements all the tested metrics.

## 1. INTRODUCTION

Understanding how specialized knowledge anchors researchers' attention, shapes boundaries, and enables interactions is a fundamental issue across various fields of science and technology studies (Gieryn, 1999). This paper focuses on a fundamental dimension of scholarly communication: research interest similarity. This concept is relevant for a wide range of issues, including technical challenges such as field normalization in scientometric indicators (Hicks, Wouters et al., 2015; Ioannidis, Boyack, & Wouters, 2016) and the measurement of interdisciplinarity (e.g., Cantone, 2024; Stirling, 2007; Wang & Schneider, 2020). It is also central to substantive research, such as global inequalities in attention allocation among researchers (Gomez, Herman, & Parigi, 2022) and the evolution of problem choice in researcher careers (Gieryn, 1978; Yu, Szymanski, & Jia, 2021; Zeng, Shen et al., 2019).

The proposed measure of similarity is based on a model of the information-seeking behavior of scientists interacting with the literature. Two papers are considered more similar if they are more likely to be retrieved during a literature search following backward and forward citations. This paper-level measure can be aggregated to higher levels, such as authors, fields, or institutions. It can also be utilized to examine the similarities of papers' references and to calculate paper-level indices. Among these possible aggregates, this paper focuses on scientist-

**Citation:** Varga, A., Kojaku, S., & Silva, F. N. (2025). Measuring research interest similarity with transition probabilities. *Quantitative Science Studies*, 6, 922–939. <https://doi.org/10.1162/qss.a.13>

**DOI:**  
<https://doi.org/10.1162/qss.a.13>

**Peer Review:**  
<https://www.webofscience.com/api/gateway/wos/peer-review/10.1162/qss.a.13>

**Supporting Information:**  
<https://doi.org/10.1162/qss.a.13>

**Received:** 2 October 2024  
**Accepted:** 8 May 2025

**Corresponding Author:**  
Attila Varga  
[attila.varga@uni-corvinus.hu](mailto:attila.varga@uni-corvinus.hu)

**Handling Editor:**  
Vincent Larivière

Copyright: © 2025 Attila Varga, Sadamori Kojaku, and Filipi Nascimento Silva. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.



level similarity, as the definition is inherently tied to authors and their efforts to update their knowledge. Additionally, we consider researchers' problem selection a pressing issue and aim to contribute conceptually and methodologically by sharing our experience in aggregating paper-level data to represent authors' research interests.

We introduce a family of indices based on random walk transition probabilities (TPs). These measures have several key properties. First, we build the concept on an explicit definition of similarity by following a simple information behavior model, providing an interpretable metric. Our approach is based on an explicit definition of similarity derived from a simple model of information behavior, making it an intuitive and interpretable metric. As the construct is rooted in observable behavior, it can be empirically studied (Wilson, 1999). This also aligns with theoretical arguments and agent-based models referencing researchers' attributes and motivations (Gieryn, 1978; Rzhetsky, Foster et al., 2015). These theoretical and empirical insights can, in turn, refine the measure to incorporate factors such as agent biases or new information-update rules. Second, our measure is fundamentally symmetric, which is a desirable quality for similarity measurement. Third, we introduce continuous similarity measures, which we argue are, in certain respects, more suitable than discrete representations for tracing the cognitive organization of scientific progress. Fourth, our measure maintains more consistent precision across different analytical scales compared to related methods. To test the scale consistency and the overall validity we conduct three experiments at different levels of aggregation: collaborations, journals, and disciplines.

In Section 2 we review past approaches and relevant computational methods to research interest similarity. Section 3 defines the concept of browsing distance and its calculation based on TPs of random walks. Section 4 introduces alternative measures. These are either computationally efficient substitutes of the original basic TP metric, or similar well-known or trivial alternatives. After describing the data in Section 5, Section 6 outlines the three validation tests for the metrics. The results of these experiments and the comparison of calculation runtimes for each method are presented in Section 7. This section also evaluates the correlation of the metrics with the nodal degree. Section 8 summarizes our key findings and insights about measuring research similarity at the author level with publication data. The paper also briefly presents a Python package to calculate all the indices discussed in this paper. More information about the package can be found in the Supplementary material (S1) and at Silva (2025).

## 2. RELATED RESEARCH

Framing the stages and interactions of evolving research agendas is often done in terms of discrete topics or problems (e.g., Gieryn, 1978; Jia, Wang, & Szymanski, 2017; Rzhetsky et al., 2015). However, we argue that when the similarity of research programs is the central concern, continuous representation is more justifiable than clustering evolving research into discrete problems. While well-defined research problems can be constructed for research administration purposes or by hindsight after a research program is ended, it is more realistic to compare two scientists' research on a gradual scale, rather than based on a rigid discrete categorization.

Previous theoretical work on research interests has used the concept of research problems to explore why and under what circumstances scientists change their research focus (Gieryn, 1978). Although the term *research problem* suggests clear-cut objectives, even this line of inquiry acknowledges the difficulty of breaking research activities into well-circumscribed problems. Gieryn (1978), in his work on problem retention and change in science, discussed the span of active research problems of a scientist at a given moment, and how these interests evolve over time. According to Gieryn, scientists usually work on a unique selection of

interrelated problems. When they transition to new areas, their new problem set remains influenced by and mixed with their previous work. Given that a research agenda is defined as a structure of multiple interrelated problems, a scientist's shifts in attention tend to be gradual. Akin to this observation, Ziman (1987), while criticizing the concept of problem choice, notes that research objectives constantly shift as work progresses. Scientific research is highly uncertain, and initial formulations of research problems can evolve radically over time. More recently, Cooper (2009) also emphasized that research problems are not clearly demarcated in ongoing research.

In recent years, research on authors has gained prominence as more name-disambiguated data have become available. Current approaches to modeling researcher attention often rely on discretized representations of topics. These topics may originate from a thesaurus or a collection of keywords (Jia et al., 2017; Leahey, Beckman, & Stanko, 2017; Rzhetsky et al., 2015; Yu et al., 2021), clustering of bibliographic couplings (Zeng et al., 2019), or topic models utilizing textual information (Amjad, Daud, & Song, 2018; Rosen-Zvi, Griffiths et al., 2012). Aside from the theoretical issue of discretizing research agendas, these approaches have several practical drawbacks. Well-curated, up-to-date classifications of research problems are often unavailable, or when they do exist, they may not be at the appropriate level of resolution. While research specialties and problems can be mapped using clustering methods, clustering involves several subjective choices that affect the final output (e.g., determining the optimal clustering resolution, filtering out smaller clusters). Additionally, the distribution of clusters is often heavily skewed, requiring careful calibration to ensure that the chosen clustering method maintains a consistent level of resolution across all clusters.

Several techniques provide continuous measures of similarity, which can be applied to bibliometric data to analyze researchers and their evolving interests. Continuous representations encode the structure of topics in a spatial embedding, where papers are positioned close to other papers on similar subjects. These embeddings can be constructed from textual content (Cohan, Feldman et al., 2020; Constantino, Kojaku et al., 2025) or citation networks (Choi & Yoon, 2022; Constantino et al., 2025; Kojaku, Yoon et al., 2021; Peng, Ke et al., 2021; Tang, Qu et al., 2015). Traditionally, continuous representations have been generated using dimensionality reduction techniques applied to keyword co-occurrence matrixes (Ahlgren & Colliander, 2009) or network adjacency matrixes (Kunegis & Lommatzsch, 2009; Qiu, Dong et al., 2018). More recently, a shift toward neural embeddings—generated by neural networks—has gained momentum (Grover & Leskovec, 2016; Hamilton, Ying, & Leskovec, 2017; Kipf & Welling, 2016; Perozzi, Al-Rfou, & Skiena, 2014). These spatial representations overcome many of the limitations of discrete methods and offer numerous analytical advantages. However, neural embeddings present a critical challenge: Due to the opaque nature of neural networks, they lack interpretability. It is often unclear what the similarity in the generated space represents or how meaningful insights can be derived from it. This lack of interpretability limits their usefulness in investigations into how knowledge structures anchor researchers' attention.

The proposed method is based on random walks on networks. As a fundamental network concept (Norris, 1997), random walks have been widely used in network analysis and have played a central role in various applications. These applications include, for example, measuring node importance (Haveliwala, 2002; Jeh & Widom, 2003; Page, Brin et al., 1999), link prediction (Liu & Lü, 2010), and detecting network communities (Delvenne, Yaliraki, & Barahona, 2010; Pons & Latapy, 2005; Rosvall & Bergstrom, 2008). Among these applications the closest to our work is recommender systems (Baluja, Seth et al., 2008; Ren, Liu et al., 2014; Shi, Kong et al., 2014). In this context, random walks are typically one component within more complex algorithms. However, aside from the technical format, recommender

systems differ from our approach in one key respect. Recommender systems are optimized to retrieve the most similar entities, whereas our approach is designed to measure the full spectrum of similarity—that is, to assess the similarity even between highly dissimilar items.

A highly relevant approach for measuring node similarities in networks, which is also based on the concept of random walk TP, is personalized PageRank (PPR; Haveliwala, 2002; Jeh & Widom, 2003). PPR extends the PageRank algorithm to capture node-specific importance scores by teleporting the walker back to its starting node. While PPR has been successfully applied in recommendation tasks (Bagci & Karagoz, 2016; Jiang, Liu et al., 2018), it is less suitable for measuring research interest similarity, for several reasons. First, this TP measure is suitable for retrieving the most similar nodes in comparison to the input node, making it less effective at assessing similarities between more distantly related nodes. In other words, it is designed to distinguish highly similar entities, but it is not sensitive to low-level similarities. Second, PPR produces asymmetric similarity scores (the similarity from node A to B differs from B to A), which contradicts the intuition that similarity should be symmetric. Third, exact computation of PPR is expensive for large networks. While approximate methods exist (e.g., Wu, Gan et al., 2021), they often trade off accuracy for computational efficiency. This trade-off arises because PPR lacks a cutoff to limit random walks, meaning that, theoretically, it can explore the network indefinitely, resulting in a fully dense similarity matrix.

Recently, recognizing the limitations of bibliographic coupling (BC) and cocitation (CC) as paper similarity measures (due to the sparsity of such links in citation networks), Yun (2022) tested the efficiency of PPR and node2vec embeddings (N2V; Grover & Leskovec, 2016; Perozzi et al., 2014) on node split networks (Yun, Ahn, & Lee, 2020) as an alternative approach. This work is highly relevant to our study, as we also investigate the same similarity measures within a bibliometric context. Yun (2022) found that many highly similar paper links could not be detected by bibliographic coupling and cocitation methods. Our study further highlights the inadequacy of BC and CC as similarity measures and shows that N2V and PPR have low correlation, agreeing mostly at high similarity ranges.

### 3. CONCEPTS AND CALCULATION

The concept of research interest similarity is based on how scientists learn about new literature related to their research agenda. A typical scientist would begin with a set of papers they are already familiar with. By following the forward or backward references of these papers, they would discover new papers. This process would then continue iteratively, as the scientist explores the references or citations of the newly found papers, until they acquire the desired new information about the current state of the art. In information retrieval, this search procedure is often referred to as *pearl-growing search* or *snowballing search*. It is an effective method for retrieving complex information (Greenhalgh & Peacock, 2005) and is likely familiar to most researchers. Based on this form of information retrieval, we define our similarity measure as follows: Two papers are more similar if, during a literature search starting from one of the papers, it is more likely to retrieve the other paper.

We do not claim that this is the primary way researchers discover new and relevant papers. However, we assume that this process, along with other mechanisms serving the same function (such as attending conferences, discussions with colleagues, and social media interactions), would yield similar results. Therefore, we consider this definition to be sufficiently general. It is important to note that we are not proposing a new information retrieval model. Rather, we are drawing upon an existing concept in information retrieval and assuming that it is most likely a typical information-seeking behavior in the sciences.

Our operational definition of similarity is based on the observation that this type of literature search resembles a walk in a citation network. Therefore, a scientist's search behavior can be modeled as a random walk. The operational definition of similarity is the following: The similarity of two papers  $i$  and  $j$  is proportional to the probability that a fixed-length random walk from  $i$  passes through  $j$ . With this definition, we introduce a generalizable similarity measure for graphs, which, in this context, applies to academic papers. The research interest similarity of authors or the similarity of scientific fields can then be derived by aggregating the similarities across sets of papers, a topic we explore in more detail in the following sections. At this point, we emphasize that our approach provides a concise conceptual framework for understanding paper similarity. Unlike previous methods, our framework is based on an explicit mechanism of information update, which is rooted in purposeful action and shaped by the enabling/constraining information structure, encoded as a citation network.

The specific measure implemented here is the probability that a random walk starting from node  $i$  transitions through a target node  $j$  at any step during the walk in  $t$  steps. The following definition assumes that the network is connected and undirected, and these assumptions are consistent with the conceptual definition above. We start the definition with  $P_{ij}^\tau$ , which is the probability that a random walk from node  $i$  is at node  $j$  after  $\tau$  steps. This TP can be calculated by summing up the probabilities of every possible  $\tau$ -step-long random walk, starting from  $i$  and terminating at  $j$ . The TP of a single random walk is given by multiplying the reciprocals of each node's degree along the walk from  $i$  to  $j$ , but *not* including the terminal node  $j$ . Finally, summing up all  $P_{ij}^\tau$  probabilities from 1 to  $t$  ( $\{1, \dots, \tau, \dots, t\}$ ) gives the probability that a random walk from  $i$  crosses  $j$ .

This probability is not symmetric ( $P_{ij}^\tau \neq P_{ji}^\tau$ ), because if, let's say,  $j$  has a higher degree than  $i$  ( $k_i < k_j$ ),  $j$  would be visited more often by random walks than  $i$ . We symmetrize the  $P_{ij}^\tau$  probabilities by dividing them with  $j$ 's degree:  $P_{ij}^\tau/k_j$ . This essentially means that the computation of a single walk's TP changes to multiplying together *all* nodes' degree reciprocals, including  $j$  as well. This operation is an addition to the conceptual definition. The rationale for this step, aside from simplifying the measure by symmetrizing it, is that the expected similarity of a paper should not be higher just because it has more references and/or citations. Although it is true that higher impact papers, or papers with more references, are retrieved more often in a search, for our purposes, it is unwanted to assign higher average similarities for highly cited papers.

The final TP, which is our basic measure  $T_{ij}^t$ , can be computed by using matrix operations in the following way. Denoted by  $A$  the adjacency matrix of the citation network, where  $A_{ij} = 1$  and  $A_{ij} = 0$  represents that a paper  $i$  is connected by an edge with paper  $j$ , or not, respectively. Recall that we ignore the direction of the citations and thus  $A_{ij} = A_{ji}$ . We use the transition matrix  $P$  of random walks, where the TP of a walker from paper  $i$  to paper  $j$  is given by  $P_{ij} = A_{ij}/k_i$ , or equivalently,  $P = D^{-1}A$ , where  $D$  is a diagonal matrix, with diagonal entries being the degree of nodes, i.e.,  $D_{ii} = k_i$ . The TP that a walker from node  $i$  would make the  $\tau + 1$ th step to node  $j$  after  $\tau$  steps can be found at the  $(i, j)$ th entry of the  $\tau + 1$ th power of the matrix  $P$ :  $(P^{\tau+1})_{ij}$ . Finally,  $T_{ij}^t$  is given by adding up each  $\tau$ th powers of the transition matrix from 1 to  $t$ , and applying the appropriate normalizations:

$$T_{ij}^t = \frac{1}{t} \left( \sum_{\tau=1}^t P^\tau D^{-1} \right)_{ij}$$

The division by  $t$  ensures that the final value correctly expresses the TP of a single step along a  $t$ -length walk; in other words, it ensures that  $T_{ij}^t$  is bound by 1.

#### 4. ALTERNATIVE FORMULATIONS

Besides the above defined basic metric  $T_{ij}^t$ , we considered six alternative indices. The first group consists of measures derived from our original concept, designed to address specific shortcomings of  $T_{ij}^t$ . The second group includes well-known measures that may appear conceptually similar to our approach. By incorporating this latter group, we aim to highlight the specific advantages of our method.

Calculating the proposed TP is computationally demanding. The TP is based on powers of  $P$  that may hold  $NN$  transition probabilities, where  $N$  is the number of nodes in the network. The number of transition probabilities increases quadratically with respect to  $N$ , and thus it is practically challenging to hold them in memory for large networks. For this reason, it is worth exploring alternative measures of  $T_{ij}^t$ . We considered two such alternatives, which are conceptually similar to the above defined TP, and would be reasonable substitutes of the original formulation.

As the first alternative, we propose a scalable method to estimate the  $T_{ij}^t$  values by simulating random walkers in the networks. Namely, we launch  $n$  random walkers from each node  $i$ , and all the walkers make  $t$  steps. The crossing frequency of these walks on node  $j$  is  $x_{ij}^t$ . The estimated transition probability is

$$ET_{ij}^t = \frac{x_{ij}^t}{ntk_j}.$$

While the matrix computation necessarily gives the  $T_{ij}^t$  of all the node pairs in the network, the advantage of  $ET_{ij}^t$  is that it enables the sampling of node pairs. However, the estimation approach has some caveats.  $ET_{ij}^t$  can be zero if there is no path with length  $t$  between the nodes (in which case  $T_{ij}^t$  is also zero), or if  $T_{ij}^t$  is so small that none of the random walkers were able to span the network structure between nodes  $i$  and  $j$  during the estimation procedure. This essentially means that the estimation has failed. One can increase the number of walkers to obtain an estimate, but it's important to bear in mind that this approach exhibits diminishing returns. The chance that at least one walker will cross and give an estimated value rises at a decreasing rate with the number of walkers delegated to this task. Also note that  $ET_{ij}^t$  is only approximately symmetric. Assuming  $k_i > k_j$  simulated random walks from  $i$  are more likely to give an estimate for  $ET_{ij}^t$  instead of remaining zero. This is because higher degree starting nodes of random walks spread the walks more evenly in the network, and these walks are more likely to reach distant nodes. Therefore, when we calculate  $ET_{ij}^t$  we simulate walks from both directions and take the average.

Another way to reduce the proportion of zero estimates and gain better precision for low similarity ranges is to increase  $t$ . Increasing  $t$  can aid the walker to span distant nodes with low  $T_{ij}^t$ . The choice for walk length  $t$  can help to balance out the problem of reachability, and the fact that, with the increase of  $t$ ,  $P^t$  is converging to a stationary distribution, which is proportional to the degree of the nodes, and would not provide meaningful information about similarity. We will return briefly to the issue of choosing the value of  $t$  in the concluding section.

The second substitute introduced here, which still adheres to the behavioral model, is a crossover of TP and shortest paths (SP), and it is referred to as the average shortest path transition probability ( $ST_{ij}^t$ ). For this measure one has to find all the shortest paths between two nodes  $i$  and  $j$ , calculate the transition probability  $p_{ij}$  of each shortest path  $p_{ij} = \prod_{\tau=1}^t \frac{1}{k_{\tau}}$  where  $k_{\tau}$  is the degree of the  $\tau$ th node on the shortest path, and average out these  $p_{ij}$  probabilities. This is a symmetric quantity; however, conceptually it detours from the original definition. One can interpret this as a knowledgeable agent's search, whose browsing behavior is informed or biased to find the most direct connection between papers  $i$  and  $j$ , and narrowing the search to corpuses of related literature, such as by considering the paper's content.

The first and most basic alternative of our methodological approach is the shortest path length. It is the length of any shortest path between two nodes and can be computed by most network analysis programs. The second alternative measure is derived from the traditional science mapping tools BC and CC. It is the Jaccard coefficient of the two papers neighborhood in the citation network (denoted as  $N(i)$ ), which includes all forward and backward references:

$$JAC = \frac{|N(i) \cap N(j)|}{|N(i) \cup N(j)|}$$

It expresses the level overlap between the two papers' forward and backward references.

We considered the N2V algorithm as an alternative of  $T_{ij}^t$ . N2V is a technique to embed nodes of a network into a vector space, generating a vector for each node. The embedding is based on a neural network trained on the sequence of nodes visited by random walks. It has been shown that the embedding generated by N2V is equivalent to the eigenvectors of the matrix  $T^t = \frac{1}{t} \sum_{\tau=1}^t P^{\tau} D^{-1}$  (Qiu et al., 2018), which relates to our measure.

Finally, we tested PPR. As was mentioned in Section 2, PPR is a widely used similarity measure that is conceptually similar to our approach, as it is also based on random walk TPs. However, PPR is not well-suited for measuring low-range similarity, due to the restart mechanism in TP calculation. At each step of a random walk, the algorithm has a teleportation probability  $p$  that allows it to return to the source node. As the first step is always taken, the probability of a walk of length  $l$  (where  $l > 1$ ) occurring decreases exponentially with  $(1 - p)^l$ . In simple terms, PPR prioritizes exploring the immediate vicinity of the focal node, whereas  $T_{ij}^t$  does not exhibit this bias. To compute PPR efficiently, we implemented the Power Iteration with Forward Push method (Wu et al., 2021).

## 5. DATA

The citation data is sourced from the Web of Science Science Citation Index (SCI), while disambiguated author information comes from the Microsoft Academic Graph (MAG), which has been matched to papers indexed in SCI. Using DOIs and bibliographic metadata, we successfully matched 90% of the papers to MAG. Research field information is derived from Web of Science subject categories (SCs). We relied on MAG for author names, as Web of Science does not provide disambiguated author identities.

For our tests, we selected three research fields plus a multidisciplinary journal set. Each field is analyzed for a focal year between 1980 and 2019. The multidisciplinary set includes all articles from *Nature* and *Science*. Table 1 summarizes the selected fields and focal years. This

**Table 1.** Summary table of data sets

Subject category	Abbreviation	Year	Nodes	Edges
Astronomy & Astrophysics	Astro	1980	65,535	983,206
Clinical Neurology	Neurol	1990	78,159	801,260
Sociology	Soc	2010	46,832	398,512
Multidisciplinary	Multi	2019	51,942	356,930

selection ensures disciplinary diversity and spans different time periods, enhancing the validity of the results. The choice of fields and specific years was primarily constrained by computational feasibility, as calculating  $T_{ij}^t$  is resource intensive.

Our goal was to create a diverse data set with manageable citation networks, allowing for efficient  $T_{ij}^t$  computation. The citation networks consist of 20-year slices of SCI-indexed papers, where the last year of each window is the focal year, and the first year is 20 years prior. For each field, the initial selection includes all articles from relevant journals within the given period. Those papers that don't have at least three connections in the network (considering all citations and references) are removed, and after that step the main component is selected for the analysis.

## 6. VALIDATION METHODS

The testing scenarios are designed to evaluate the performance of the similarity metrics across diverse use cases. Our primary concern is whether these metrics can consistently measure similarity at different scales and levels of aggregation. To devise these tests, we leverage the hierarchical nature of scientific research, where specific research goals are nested within specialties, which in turn belong to broader disciplines. These levels of scientific organization serve as ground truth for our validation tests.

Before detailing the validation tests, we briefly address why a consistent measurement scale is particularly important in science studies, whereas it is less critical in recommender systems and information retrieval. Estimating and distinguishing similarity at the mid- and macroscales is crucial in several research contexts, particularly when studying information diffusion beyond local and direct interactions (Nakamura, Suzuki et al., 2011; Varga, 2019). It is also relevant in interdisciplinarity research, where long-distance interactions play a significant role (Larivière, Haustein, & Börner, 2015). Furthermore, it is essential for studies examining the relationship between cognitive structures and institutional processes within broader disciplinary organizations and policies (Becher & Trowler, 2001; Gieryn, 1999; Mohr & White, 2008). Scientists frequently engage in collective decision-making where research interest similarities are low but still influential in forming a collective identity. Such scenarios commonly arise in faculty hiring, funding allocation, and large-scale data collection efforts (Brooks, Heidorn et al., 2016).

To evaluate the validity and precision of the similarity metrics, we designed three levels of validation tests. At the microlevel of coauthorship evolution, we tested how well a given method predicts future first-time collaborations based on past research similarity. At the midlevel of research specialties—approximated by journals—we measured how accurately the method distinguish between papers published in the same journal versus papers from different journals within the same discipline. Finally, at the macrolevel, the test was to classify

papers based on disciplinary background. Our expectation is that N2V will better capture local relationships but provide a coarser representation at the mid-to-macro range, as N2V is known to be less accurate for weaker similarity ranges (Grover & Leskovec, 2016). In all three cases, the test data consist of two variables. The first is a binary variable indicating whether authors have collaborated, papers are in the same journal, or papers belong to the same discipline. The second variable is a predictor/classifier variable, which is the tested similarity metric.

In the following, we will describe the collaboration prediction task in detail. This scenario also introduces practical methods for measuring author pair similarity. The collaboration prediction test consists of two sets of author pairs: future collaborators (authors who have never collaborated before but will collaborate in the year following the focal year) and noncollaborators (randomly matched authors who did not collaborate before and will not collaborate in the following year). All selected authors are actively publishing and collaborating in the following 3 years after the focal year with someone. These selection criteria are necessary to keep the two sets comparable. Both the collaborator and noncollaborator author pair sets are based on a sample of 500 authors, except in sociology (232 authors) and multidisciplinary journals (431 authors) due to data constraints. Each author is paired with one of their future collaborators for the true collaboration set and with a random author for the noncollaborator set. We use a 5-year window to select papers written by the focal author, which serves as the basis for computing potential collaborator similarity. This same window is used to determine whether authors have previously collaborated. To measure research interest similarity, we consider only papers where the focal author is listed as the first or last author, assuming these roles better reflect their research interests. Our experience suggests that this selection improves prediction accuracy. The final author similarity score is the average pairwise similarity of their selected papers.

For the journal classification task, every journal within the given discipline was first selected for analysis. Then, 20 papers were randomly chosen from each journal. In the case of multidisciplinary data, where only two journals were present, 500 papers were selected from each. Each paper was then matched with one from the same journal and another from a randomly chosen journal. Journals that did not have a sufficient number of papers within the given period to meet these sampling criteria were excluded from the analysis.

For the macrorange test, we classified the subject categories (SCs) and their corresponding journals into six broadly defined disciplines: environmental sciences, agriculture, medicine, biology, chemistry, and physics. This test was conducted exclusively on multidisciplinary journal data (papers from *Nature* and *Science*). A paper's disciplinary background was determined by identifying the most frequently occurring discipline among its references. The sample consisted of 10,000 paper pairs, with half belonging to the same discipline and the other half to different disciplines. To ensure balanced representation, all disciplines were given equal weight in the data set.

## 7. RESULTS

We begin with descriptive statistics. These exploratory results are based on random samples of dyads with a size of 2,000 paper pairs in each field. Table 2 lists the tested measures and their abbreviations, while the Supplementary material (S2) presents the parameterization of the measures. The proportion of paper dyads with at least one overlapping citation or reference is less than 1% in all cases. Therefore, we do not present descriptive and bivariate statistics for JAC (Jaccard similarity) in the following. Figure 1 displays the distribution of the measures across the four dyad samples. The distribution of the TP measures roughly follows a lognormal

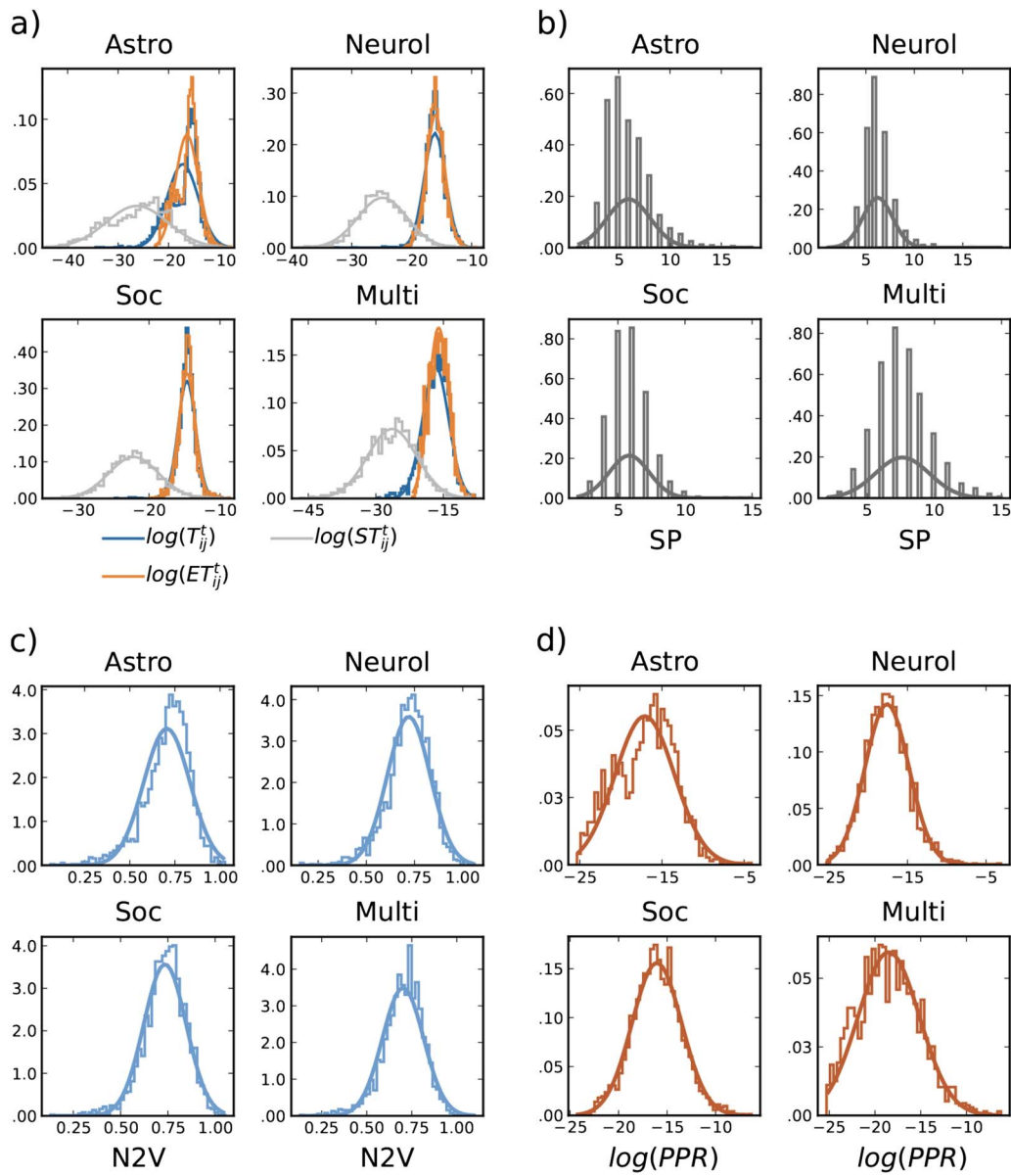
**Table 2.** Tested metrics and their abbreviations

Abbreviation	Definition
$T_{ij}^t$	TP calculated using matrix operation
$ET_{ij}^t$	TP calculated using simulated random walks
$ST_{ij}^t$	Average shortest path TP
SP	Shortest path distance
N2V	Cosine similarity of nodes in embedding space generated with node2vec
PPR	Personalized page rank
JAC	Jaccard similarity of backward and forward citations

distribution (Figures 1(a) and (d)). The exception to this lognormal shape is Astronomy & Astrophysics, which exhibits a bimodal lognormal distribution. We attribute this uniqueness to the discipline's pronounced dual organization into two subfields: cosmology and observational astronomy (Varga, 2018). Within-subfield paper pairs are similar to each other, while between-subfield pairs are dissimilar, creating two distinct peaks.  $T_{ij}^t$  and its estimate  $ET_{ij}^t$  are almost identically distributed. The N2V measure of similarity, represented by the cosine distance in the vector space (Figure 1(c)), is very close to being normally distributed in all cases. Finally, PPR, similar to the other TP measures, has a lognormal distribution with two peaks in the case of Astronomy & Astrophysics (Figure 1(d)).

The direct implementation of the measurement concept is  $T_{ij}^t$  and for this reason we examined how it relates to all the other similarity indices (Figure 2). Overall, its direct estimate  $ET_{ij}^t$  has the strongest correlation with  $T_{ij}^t$  (Figure 2(a)). The next strongest association of  $T_{ij}^t$  is with  $ST_{ij}^t$  and PPR (Figures 2(a) and (d)). The relationship of  $T_{ij}^t$  and  $ST_{ij}^t$  is not perfectly linear, but  $ST_{ij}^t$  gives consistently good approximations for all ranges of  $T_{ij}^t$  and in all fields. SP has a weaker correlation with  $T_{ij}^t$  (Figure 2(b)) compared to the previous indices. As we mentioned above, we expected that N2V will correlate less with  $T_{ij}^t$  at weaker similarity ranges. We fitted piecewise linear regression models (OLS) to see if N2V and  $T_{ij}^t$  has differing magnitudes of association at higher and lower TP values (Figure 2(c)). The regression lines were fitted above and under the median  $T_{ij}^t$ . These slopes are considerably different; for high  $T_{ij}^t$  values, the association is much stronger, and in most cases the association turns negative at weaker similarities. We produced similar results as Yun (2022) in relation to N2V and PPR. We explored the parameters of the embedding method to make sure that the nonlinear and uneven association with  $T_{ij}^t$  is not an idiosyncratic result. Our analysis can be found in the Supplementary material (S3), and we can conclude that this discrepancy is observable with all parameter configurations for N2V.

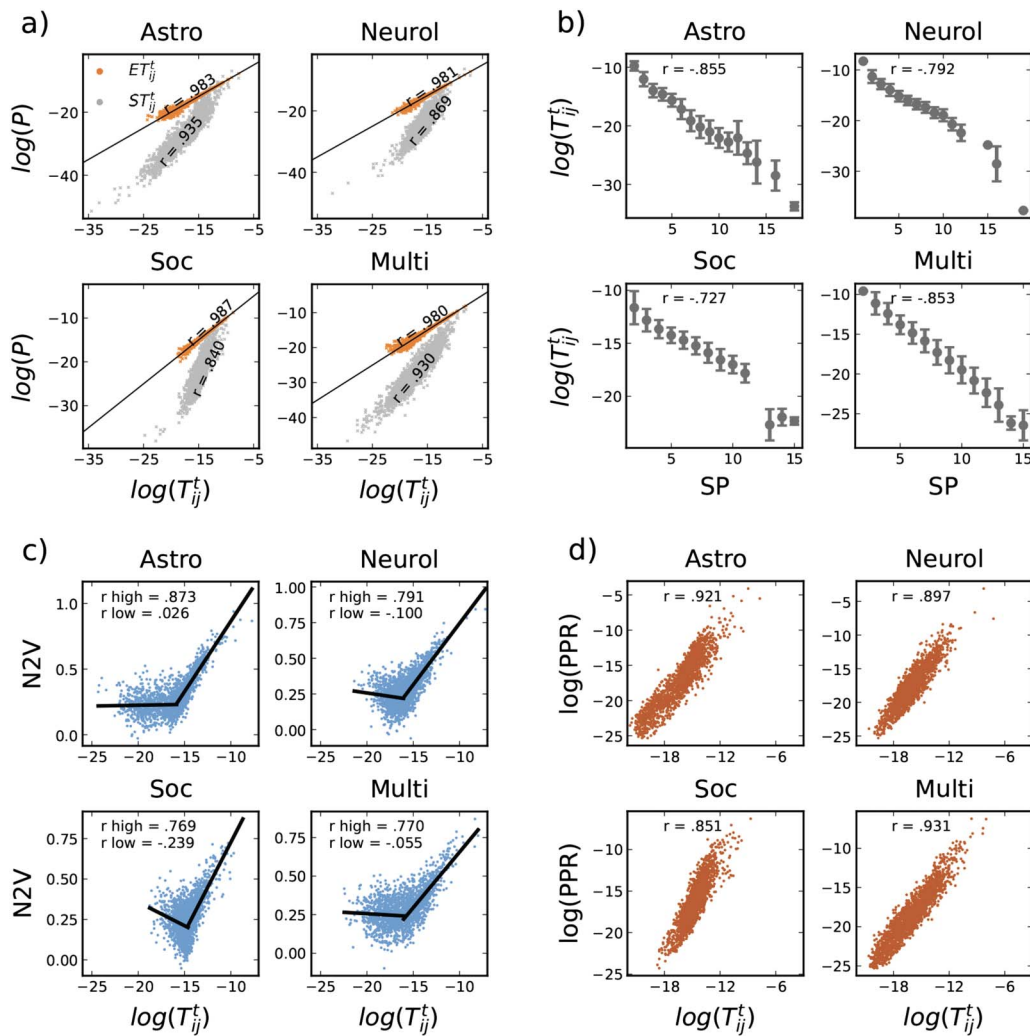
To compare the classification performance of the measures, we used the area under the curve score (AUC). Looking first at the coauthorship prediction (Figure 3), we can observe that, on average, the TP measures perform quite well, with AUC values around 0.85–0.9 in the best cases. However, in the case of clinical neurology, the average AUC drops to 0.74. Perhaps in this field, the clinical work setting mixes a diverse set of researchers, and the clinical setting may be a more important determinant of collaborative work than one's actual research specialty. However, this is merely speculation. Returning to the measurement evaluation, JAC is the worst performer, while SP is the second worst.



**Figure 1.** Distribution of the measures with fitted normal curves. (a)  $T_{ij}^t$ ,  $ET_{ij}^t$ , and  $ST_{ij}^t$ ; (b) SP; (c) N2V cosine distance; (d) PPR.

Figure 4 presents the results for journal classification. The average AUC score drops from coauthor prediction, and the performance spread is wider. The scores for the multidisciplinary data set are at 0.5, suggesting that it is difficult to differentiate papers published in *Science* and *Nature*. The best performers are  $T_{ij}^t$  and  $ET_{ij}^t$  (0.73–0.79 AUC), followed by N2V (0.70–0.77 AUC). PPR significantly falls behind N2V in the case of neurology and sociology (0.66–0.70 AUC). SP and JAC are the worst performers, with JAC having almost no predictive capability, because at that similarity range reference and citation overlap is down to 1–2%.

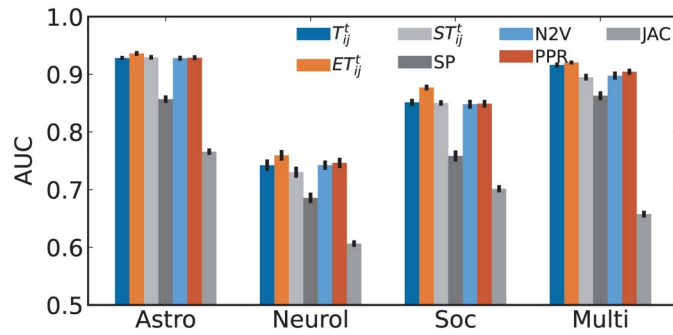
The classification of disciplinary affiliations provides further insight into the utility of the studied measures (Figure 5(a)). The main finding is that N2V significantly underperformed compared to all TP measures, including PPR. While the best performers  $T_{ij}^t$  and  $ET_{ij}^t$  scored around 0.73, the AUC of N2V is 0.6. PPR and  $ST_{ij}^t$  performed similarly to  $T_{ij}^t$  and  $ET_{ij}^t$  with



**Figure 2.** Bivariate distribution of  $T_{ij}^t$  and the alternative measures.  $r$  stands for Pearson correlation. (a)  $T_{ij}^t$ ,  $ET_{ij}^t$ , and  $ST_{ij}^t$ ; (b) SP, means, and standard deviations; (c) N2V cosine distance. The piecewise regression lines are fitted under and above the median of  $T_{ij}^t$ . (d) PPR.

AUC scores 0.71. SP came closest to the TP measures in this test, with a 0.67 AUC score. JAC is omitted from the plot, as almost all paper pair values were zero.

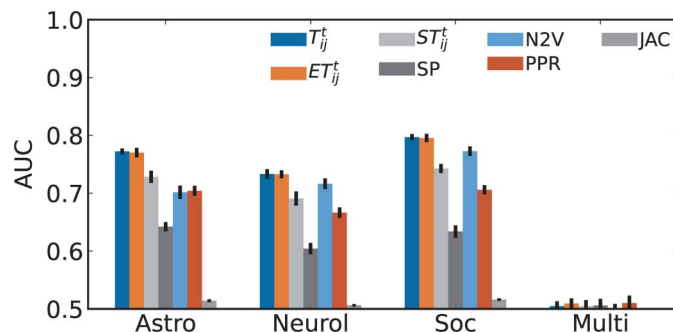
We also present the results of an additional analysis of the macro disciplinary structure, specifically focusing on the performance of TP measures versus N2V. The purpose is to show further evidence that N2V underperforms at the macro level. We arrange the disciplines from life to physical sciences, starting with environmental sciences and ending with physics. Figures 5(b) and (c) show similarity measures for each discipline in a matrix format for  $T_{ij}^t$  and N2V. Based on the logic of the ordering, we expected that the similarities would be strongest around the diagonal of the resulting matrix, as any given discipline should be closest to its neighboring discipline along the continuum. This pattern is more noticeable in the case of  $T_{ij}^t$ , where the diagonal is more prominent. Agriculture, medicine, biology, and chemistry show an affinity with one another. In both figures, biology is relatively strongly related to all life sciences, though this is more evident in the case of  $T_{ij}^t$ .



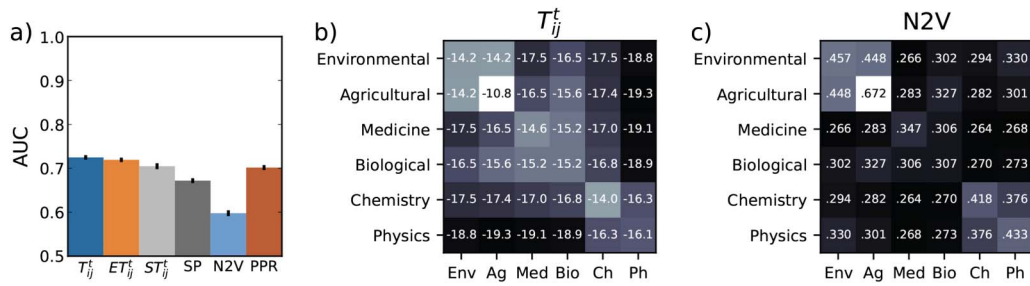
**Figure 3.** Coauthor prediction AUC scores. Error bars are 95% confidence intervals from bootstrapping.

In Section 3 we mentioned that the TP measures are related to the nodal degree. A random walk is more likely to pass through a high degree node. We evaluated this degree bias of the tested measures and show the results on Figure 6. These results are based on the randomly sampled paper pairs. Because PPR is asymmetric, we plotted the correlation coefficient for the nodes in the pair separately for each measure. N2V has a negligible correlation with degree. This desirable property of N2V is well known. JAC also has a small degree bias according to our test. The Spearman correlation of  $T_{ij}^t$  and  $ET_{ij}^t$  with degree is moderate (0.05–0.19).  $ST_{ij}^t$  has a slightly higher correlation, ranging between 0.09 and 0.21. The degree bias of PPR has the same range (0.09 to 0.21) on the source node side of the dyad. However, the target node’s degree has a strong correlation with PPR. This is not a surprise given that PageRank was introduced originally as a webpage popularity ranking algorithm. Finally, shortest paths are also strongly dependent on the degree.

The evaluation of the runtimes is shown in Figure 7. These tests were conducted on random networks generated using the stochastic block model (Karrer & Newman, 2011). Such networks replicate the underlying modular structure commonly observed in real-world networks, where nodes are grouped into communities (such as disciplines), with a higher probability of connections within the same group and a lower probability between different groups. This approach allows us to evaluate the performance of our methods under controlled conditions, simulating community structures found in scholarly networks. Our results indicate that the speed for  $ET_{ij}^t$  scales best with network size, with  $T_{ij}^t$  feasible only for small networks in practice. Overall, PPR calculated with the Power Iteration with Forward Push method appears to be the slowest. However, the computational speed somewhat stabilizes for larger networks.



**Figure 4.** Journal classification AUC scores. Error bars are 95% confidence intervals from bootstrapping.



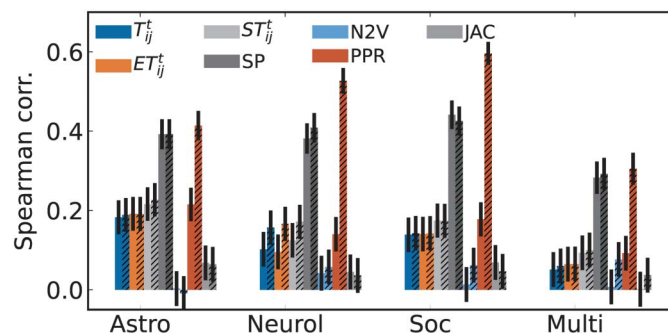
**Figure 5.** Macro structure prediction and mapping. (a) Disciplinary coclassification AUC scores. Error bars are 95% confidence intervals from bootstrapping. (b) and (c) Discipline similarity matrixes. The values in the cells are averages of the pertaining similarity measures. The shading is proportional to cell values.

The employed parameter setting contributes to the method’s slowness. This parameter setting (S2) is necessary to achieve a sufficient level of precision for the validation tests. These tests were demanding for PPR because the probability of the walker restarting had to be set very low, so that the walkers could reach beyond the immediate neighborhoods of the focal nodes.

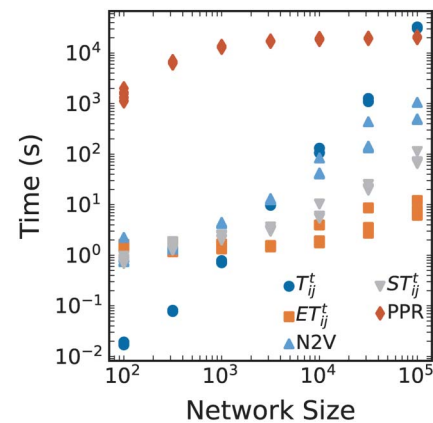
### 8. CONCLUSIONS

We have presented an interpretable, continuous, and symmetrical family of measures for paper and author similarity. The definition of these measures has explicit conceptual foundations, which leads straightforwardly to their operational definition based on random walk TPs. We found that our main metric  $T_{ij}^t$  and its direct estimate  $ET_{ij}^t$  are the most consistent measures of similarity across different scales of analysis (coauthorships, journals, and disciplines), outperforming our other proposed solutions as well as rival similarity measures. Additionally,  $ET_{ij}^t$  was found to be reasonably lightweight to implement. The closest alternative to the concept behind  $T_{ij}^t$  and  $ET_{ij}^t$  is PPR, which prioritize local similarity precision over long distance similarity exploration. This explains why it performed slightly less effectively despite considerable computational investment.

The major limitation of our approach stems from the fact that publication data does not represent equally well the research interest of scientists across different fields of inquiry. Moreover, our metrics exhibit some level of degree bias: High-degree nodes (papers) appear to be more similar to other nodes in the citation network. This is an undesirable quality for a



**Figure 6.** Degree bias of the seven similarity measures. Error bars are 95% confidence intervals of the Spearman correlation coefficient. The plain bars represent the source nodes, and the patterned bars indicate the target node’s correlation coefficient.



**Figure 7.** Runtime benchmarks on random networks. Ten random networks for each network size and method.

similarity measure. However, we want to point out that by moving higher degree papers closer,  $T_{ij}^t$  remains consistent with its conceptual definition: High degree papers are simply easy to reach while someone is browsing the literature. In other words, highly cited papers are more visible and more likely to bridge authors and fields. Finally, we want to avoid the impression that we aim to measure the actual interests of researchers. Rather, our terminology expresses the interests of a generic information-searching agent without a concern about the specific content of the browsed literature. The logic of this approach is to map the enabling and constraining structure of information encoded in the citation network to trace similarity. However, as we expressed in the introduction, the measurement approach can be developed further to include nuances that represent more closely some specific search profiles.

Choosing the appropriate walk length for  $T_{ij}^t$  and  $ET_{ij}^t$  is an important step. If the network is very sparse and/or highly clustered, zero similarity values could be an issue. In this case, one should increase the walk length. Given the small-world property of most networks, the shortest path length between two nodes should be relatively short. A relevant property of TP—related to the previous discussion on degree correlation—is that as the walk length  $\tau$  increases, the TP defined as  $(P^\tau)_{ij}$  converges to the degree of  $j$ . However,  $\tau$  must be very large to induce this effect in large networks. Overall, we did not find a reason to go beyond walk lengths over approximately 20, as this would undesirably slow down computation.

We have two observations about relying on paper-level information to represent author-level behavior. Both observations highlight a general issue stemming from the loose relationship between authorship status and author contributions. These issues become more pronounced as the prevalence of multiple authors increases. First, in our experience, only considering papers where the focal author is in the first or last authorship position serves as a better predictor of future collaborators. We used this approach across all disciplines, but it may be beneficial to calibrate this step for certain disciplinary publishing cultures. The main disadvantage of this choice is information loss, as it becomes harder to assess an author’s interests if they have not published enough papers as a first or last author. The second observation is that, in some fields, publications represent researchers’ true interests less accurately. As we already mentioned, the usefulness of paper-level data to represent author interests is inherently limited. This issue was particularly evident in clinical neurology, where we speculated that the work setting of clinical research might explain the discrepancy. It is possible that substantive

research interests and clinical practice do not always align when researchers collaborate. Perhaps data collection and specialized expertise diverge in clinical research, although both contribution types qualify for authorship.

#### AUTHOR CONTRIBUTIONS

Attila Varga: Conceptualization, Formal analysis, Methodology, Software, Validation, Writing—original draft, Writing—review & editing. Sadamori Kojaku: Methodology, Software, Writing—review & editing. Filipi Nascimento Silva: Formal analysis, Methodology, Software, Writing—review & editing.

#### COMPETING INTERESTS

The authors have no competing interests.

#### FUNDING INFORMATION

The authors thank the support from the Air Force Office of Scientific Research under Award #FA9550-19-1-0391. This work used JetStream2 at Indiana University through allocation CIS230183 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

#### DATA AVAILABILITY

Web of Science data are the property of Clarivate. Disambiguated author names are available at [openalex.org](https://openalex.org).

#### REFERENCES

- Ahlgren, P., & Colliander, C. (2009). Document–document similarity approaches and science mapping: Experimental comparison of five approaches. *Journal of Informetrics*, 3(1), 49–63. <https://doi.org/10.1016/j.joi.2008.11.003>
- Amjad, T., Daud, A., & Song, M. (2018). Measuring the impact of topic drift in scholarly networks. In *Companion Proceedings of The Web Conference 2018* (pp. 373–378). <https://doi.org/10.1145/3184558.3186358>
- Bagci, H., & Karagoz, P. (2016). Context-aware friend recommendation for location based social networks using random walk. In *Proceedings of the 25th International Conference Companion on World Wide Web* (pp. 531–536). <https://doi.org/10.1145/2872518.2890466>
- Baluja, S., Seth, R., Sivakumar, D., Jing, Y., Yagnik, J., ... Aly, M. (2008). Video suggestion and discovery for YouTube: Taking random walks through the view graph. In *Proceedings of the 17th International Conference on World Wide Web* (pp. 895–904). <https://doi.org/10.1145/1367497.1367618>
- Becher, T., & Trowler, P. (2001). *Academic tribes and territories* (2nd ed.). SRHE and Open University Press.
- Brooks, C. F., Heidorn, P. B., Stahlman, G. R., & Chong, S. S. (2016). Working beyond the confines of academic discipline to resolve a real-world problem: A community of scientists discussing long-tail data in the cloud. *First Monday*, 21(2). <https://doi.org/10.5210/fm.v21i2.6103>
- Cantone, G. G. (2024). How to measure interdisciplinary research? A systemic design for the model of measurement. *Scientometrics*, 129, 4937–4982. <https://doi.org/10.1007/s11192-024-05085-1>
- Choi, J., & Yoon, J. (2022). Measuring knowledge exploration distance at the patent level: Application of network embedding and citation analysis. *Journal of Informetrics*, 16(2), 101286. <https://doi.org/10.1016/j.joi.2022.101286>
- Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. (2020). SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 2270–2282). <https://doi.org/10.18653/v1/2020.acl-main.207>
- Constantino, I., Kojaku, S., Fortunato, S., & Ahn, Y.-Y. (2025). Representing the disciplinary structure of physics: A comparative evaluation of graph and text embedding methods. *Quantitative Science Studies*, 6, 263–280. [https://doi.org/10.1162/qss\\_a\\_00349](https://doi.org/10.1162/qss_a_00349)
- Cooper, M. H. (2009). Commercialization of the university and problem choice by academic biological scientists. *Science, Technology, & Human Values*, 34(5), 629–653. <https://doi.org/10.1177/0162243908329379>
- Delvenne, J.-C., Yaliraki, S. N., & Barahona, M. (2010). Stability of graph communities across time scales. *Proceedings of the*

- National Academy of Sciences, 107(29), 12755–12760. <https://doi.org/10.1073/pnas.0903215107>, PubMed: 20615936
- Gieryn, T. F. (1978). Problem retention and problem change in science. *Sociological Inquiry*, 48(3–4), 96–115. <https://doi.org/10.1111/j.1475-682X.1978.tb00820.x>
- Gieryn, T. F. (1999). *Cultural boundaries of science: Credibility on the line*. Chicago, IL: University of Chicago Press. <https://doi.org/10.7208/chicago/9780226824420.001.0001>
- Gomez, C. J., Herman, A. C., & Parigi, P. (2022). Leading countries in global science increasingly receive more citations than other countries doing similar research. *Nature Human Behaviour*, 6(7), 919–929. <https://doi.org/10.1038/s41562-022-01351-5>, PubMed: 35637294
- Greenhalgh, T., & Peacock, R. (2005). Effectiveness and efficiency of search methods in systematic reviews of complex evidence: Audit of primary sources. *BMJ*, 331(7524), 1064–1065. <https://doi.org/10.1136/bmj.38636.593461.68>, PubMed: 16230312
- Grover, A., & Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 855–864). <https://doi.org/10.1145/2939672.2939754>
- Hamilton, W., Ying, R., & Leskovec, J. (2017). Inductive representation learning on large graphs. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, & R. Fergus (Eds.), *Advances in neural information processing systems* (pp. 1025–1035). Curran Associates, Inc. [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/5dd9db5e033da9c6fb5ba83c7a7e9bea9-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/5dd9db5e033da9c6fb5ba83c7a7e9bea9-Paper.pdf)
- Haveliwala, T. H. (2002). Topic-sensitive PageRank. In *Proceedings of the 11th International Conference on World Wide Web* (pp. 517–526). <https://doi.org/10.1145/511446.511513>
- Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, 520(7548), 429–431. <https://doi.org/10.1038/520429a>, PubMed: 25903611
- Ioannidis, J. P. A., Boyack, K., & Wouters, P. F. (2016). Citation metrics: A primer on how (not) to normalize. *PLOS Biology*, 14(9), e1002542. <https://doi.org/10.1371/journal.pbio.1002542>, PubMed: 27599158
- Jeh, G., & Widom, J. (2003). Scaling personalized web search. In *Proceedings of the 12th International Conference on World Wide Web* (pp. 271–279). <https://doi.org/10.1145/775152.775191>
- Jia, T., Wang, D., & Szymanski, B. K. (2017). Quantifying patterns of research-interest evolution. *Nature Human Behaviour*, 1, 0078. <https://doi.org/10.1038/s41562-017-0078>
- Jiang, Z., Liu, H., Fu, B., Wu, Z., & Zhang, T. (2018). Recommendation in heterogeneous information networks based on generalized random walk model and Bayesian personalized ranking. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (pp. 288–296). <https://doi.org/10.1145/3159652.3159715>
- Karrer, B., & Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E*, 83, 016107. <https://doi.org/10.1103/PhysRevE.83.016107>, PubMed: 21405744
- Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks (Version 4). *arXiv*. <https://doi.org/10.48550/ARXIV.1609.02907>
- Kojaku, S., Yoon, J., Constantino, I., & Ahn, Y.-Y. (2021). Residual2Vec: Debiasing graph embedding with random graphs. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems* (pp. 24150–24163). Curran Associates, Inc. [https://proceedings.neurips.cc/paper\\_files/paper/2021/file/ca9541826e97c4530b07dda2eba0e013-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2021/file/ca9541826e97c4530b07dda2eba0e013-Paper.pdf)
- Kunegis, J., & Lommatzsch, A. (2009). Learning spectral graph transformations for link prediction. In *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 561–568). <https://doi.org/10.1145/1553374.1553447>
- Larivière, V., Haustein, S., & Börner, K. (2015). Long-distance interdisciplinarity leads to higher scientific impact. *PLOS ONE*, 10(3), e0122565. <https://doi.org/10.1371/journal.pone.0122565>, PubMed: 25822658
- Leahey, E., Beckman, C. M., & Stanko, T. L. (2017). Prominent but less productive: The impact of interdisciplinarity on scientists' research. *Administrative Science Quarterly*, 62(1), 105–139. <https://doi.org/10.1177/0001839216665364>
- Liu, W., & Lü, L. (2010). Link prediction based on local random walk. *Europhysics Letters*, 89(5), 58007. <https://doi.org/10.1209/0295-5075/89/58007>
- Mohr, J. W., & White, H. C. (2008). How to model an institution. *Theory and Society*, 37(5), 485–512. <https://doi.org/10.1007/s11186-008-9066-0>
- Nakamura, H., Suzuki, S., Hironori, T., Kajikawa, Y., & Sakata, I. (2011). Citation lag analysis in supply chain research. *Scientometrics*, 87(2), 221–232. <https://doi.org/10.1007/s11192-011-0341-x>
- Norris, J. R. (1997). *Markov chains*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511810633>
- Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web*. Stanford InfoLab. <https://ilpubs.stanford.edu:8090/422/>
- Peng, H., Ke, Q., Budak, C., Romero, D. M., & Ahn, Y.-Y. (2021). Neural embeddings of scholarly periodicals reveal complex disciplinary organizations. *Science Advances*, 7(17), eabb9004. <https://doi.org/10.1126/sciadv.abb9004>, PubMed: 33893092
- Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). DeepWalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 701–710). <https://doi.org/10.1145/2623330.2623732>
- Pons, P., & Latapy, M. (2005). Computing communities in large networks using random walks. In P. Yolum, T. Güngör, F. Gürgen, & C. Özturan (Eds.), *Computer and Information Sciences—ISCIS 2005* (pp. 284–293). Berlin, Heidelberg: Springer. [https://doi.org/10.1007/11569596\\_31](https://doi.org/10.1007/11569596_31)
- Qiu, J., Dong, Y., Ma, H., Li, J., Wang, K., & Tang, J. (2018). Network embedding as matrix factorization: Unifying DeepWalk, LINE, PTE, and node2vec. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (pp. 459–467). <https://doi.org/10.1145/3159652.3159706>
- Ren, X., Liu, J., Yu, X., Khandelwal, U., Gu, Q., ... Han, J. (2014). ClusCite: Effective citation recommendation by information network-based clustering. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 821–830). <https://doi.org/10.1145/2623330.2623630>
- Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2012). The author-topic model for authors and documents. *arXiv*. <https://doi.org/10.48550/arXiv.1207.4169>
- Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4), 1118–1123. <https://doi.org/10.1073/pnas.0706851105>, PubMed: 18216267
- Rzhetsky, A., Foster, J. G., Foster, I. T., & Evans, J. A. (2015). Choosing experiments to accelerate collective discovery.

- Proceedings of the National Academy of Sciences*, 112(47), 14569–14574. <https://doi.org/10.1073/pnas.1509757112>, PubMed: 26554009
- Shi, C., Kong, X., Huang, Y., Yu, P. S., & Wu, B. (2014). HeteSim: A general framework for relevance measure in heterogeneous networks. *IEEE Transactions on Knowledge and Data Engineering*, 26(10), 2479–2492. <https://doi.org/10.1109/TKDE.2013.2297920>
- Silva, F. N. (2025). TP Similarity. *GitHub*. <https://github.com/filipinascimento/tps similarity>
- Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, 4(15), 707–719. <https://doi.org/10.1098/rsif.2007.0213>, PubMed: 17327202
- Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., & Mei, Q. (2015). LINE: Large-scale Information Network Embedding. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 1067–1077). <https://doi.org/10.1145/2736277.2741093>
- Varga, A. (2018). Novelty and foreseeing research trends: The case of astrophysics and astronomy. *Astrophysical Journal Supplement Series*, 236(1), 21. <https://doi.org/10.3847/1538-4365/aab765>
- Varga, A. (2019). Shorter distances between papers over time are due to more cross-field references and increased citation rate to higher-impact papers. *Proceedings of the National Academy of Sciences*, 116(44), 22094–22099. <https://doi.org/10.1073/pnas.1905819116>, PubMed: 31611374
- Wang, Q., & Schneider, J. W. (2020). Consistency and validity of interdisciplinary measures. *Quantitative Science Studies*, 1(1), 239–263. [https://doi.org/10.1162/qss\\_a\\_00011](https://doi.org/10.1162/qss_a_00011)
- Wilson, T. D. (1999). Models in information behaviour research. *Journal of Documentation*, 55(3), 249–270. <https://doi.org/10.1108/EUM00000000007145>
- Wu, H., Gan, J., Wei, Z., & Zhang, R. (2021). Unifying the global and local approaches: An efficient power iteration with forward push. In *Proceedings of the 2021 International Conference on Management of Data* (pp. 1996–2008). <https://doi.org/10.1145/3448016.3457298>
- Yu, X., Szymanski, B. K., & Jia, T. (2021). Become a better you: Correlation between the change of research direction and the change of scientific performance. *Journal of Informetrics*, 15(3), 101193. <https://doi.org/10.1016/j.joi.2021.101193>
- Yun, J. (2022). Generalization of bibliographic coupling and co-citation using the node split network. *Journal of Informetrics*, 16(2), 101291. <https://doi.org/10.1016/j.joi.2022.101291>
- Yun, J., Ahn, S., & Lee, J. Y. (2020). Return to basics: Clustering of scientific literature using structural information. *Journal of Informetrics*, 14(4), 101099. <https://doi.org/10.1016/j.joi.2020.101099>
- Zeng, A., Shen, Z., Zhou, J., Fan, Y., Di, Z., ... Havlin, S. (2019). Increasing trend of scientists to switch between topics. *Nature Communications*, 10, 3439. <https://doi.org/10.1038/s41467-019-11401-8>, PubMed: 31366884
- Ziman, J. M. (1987). The problem of “problem choice.” *Minerva*, 25(1–2), 92–106. <https://doi.org/10.1007/BF01096858>