CPB Netherlands Bureau for Economic Policy Analysis

Temporal Patterns in Economics Research

We study the duration of topics in economics research by looking at how much time passes between publication of textually similar papers. Using the corpus of abstracts of economics papers, as available from the RePEc dataset, we find that most papers match to papers from the same year, indicating strong common trends in the economics literature.

Nevertheless, matches as long as 14 years apart are statistically significant, suggesting there are topics that last as long. Finally, the average duration of a match has dropped from around 4 years during 1990–2005 to about 1 year starting in 2010.

CPB Discussion Paper

Andrei Dubovik, Clemens Fiedler, Alexei Parakhonyak

September 2022

Doi: https://doi.org/10.34932/29xk-nn43

Temporal Patterns in Economics Research*

Andrei Dubovik†[1], Clemens Fiedler[1], and Alexei Parakhonyak[2]

[1]CPB Netherlands Bureau for Economic Policy Analysis
[2]University of Oxford

September 16, 2022

Abstract

We study the duration of topics in economics research by looking at how much time passes between publication of textually similar papers. Using the corpus of abstracts of economics papers, as available from the RePEc dataset, we find that most papers match to papers from the same year, indicating strong common trends in the economics literature. Nevertheless, matches as long as 14 years apart are statistically significant, suggesting there are topics that last as long. Finally, the average duration of a match has dropped from around 4 years during 1990–2005 to about 1 year starting in 2010.

Key Words: RePEc, economics literature, text analysis.

JEL Classification: A14, B20, Z13.

1 Introduction

Economics as a field is ever-evolving: new methods are adopted, new questions are explored, and pressing current issues get studied. In turn, the language follows up. In this paper we attempt to quantify the speed of these changes using methods of text analysis. We are interested in how much time elapses between publishing a (working) paper and the appearance of papers that are sufficiently similar to it, as well how this time has changed over the years.

Our text analysis is based on abstracts of economics papers as found in the RePEc database. After filtering, we arrive at 420,128 working papers

*The authors thank Piek Vossen for discussion and valuable suggestions.

†Corresponding author, andrei@dubovik.eu.

written in English during 1990–2020. Focusing on working rather than published papers allows us to avoid the prevalent problem of the publication lag influencing our results. We also re-run our analysis for papers published in top-20 economics journals, according to Combes and Linnemer (2010) ranking, and arrive to qualitatively similar results. We look at how frequently various stems appear in the abstracts of each of these papers and compute the term frequency–inverse document frequency (TF-IDF) vectors for those papers. Then, using cosine similarity, we find for each paper the five[1] most similar papers published, whether in the current, consequent, or preceding years.

We ask the question, how far apart in terms of publication years are those neighbouring papers positioned? Our findings are as follows. Firstly, most of similar papers are published within one year. This result is persistent across all publication years and suggests that economists often independently work on similar topics, whether prompted by major events in the global economy or methodological advances. Secondly, the average duration between the publication of similar papers was around four years in 1990–2005, but dropped down to just one year in 2010–2020. At the same point in time, around 2007, a similar drop happened for papers published in top-20 journals, where the average duration dropped from four to two years. Thirdly, some topics stay popular for up to fifteen years, in the sense that for the period of up to fifteen years similar papers appear more frequently than what would be consistent with the uniform distribution of similar papers over the years.

Our paper makes a contribution to the literature on publishing in economics. This literature has focused on various aspects of the problem, e.g. the role of top-5 journals (Card and DellaVigna, 2013), the editorial impact on top-5 publication decisions (Önder et al., 2018), the impact of author demographics (Hamermesh, 2013) and geography (Fontana et al., 2019) on publishing, the exploration of what journals are more important for a specific field (Bellas and Kosnik, 2019). While our paper looks at the importance of narrow research questions within each field, most of the literature tends to focus on the evolution of large fields and styles (e.g., Card and DellaVigna, 2013; Angrist et al., 2017, 2020). This difference in research question determines the difference in methods: while the aforementioned papers rely mainly on JEL-codes for field classification or use manual classification of papers in terms of style, we rely on text analysis to find similar papers independently of their JEL-codes. This allows us to pick field-style combinations directly, say by the presence of terms “dataset” or “evidence” in empirical papers, or terms “monetary” or “inflation” in macroeconomics papers. Furthermore, our approach is not prone to the critique that JEL-codes might be chosen strategically (Cherrier, 2017; Önder et al., 2018).

[1] Our analysis for 1 and 10 closest papers produces similar results.

JEL-codes are suitable for the analysis of evolution of larger fields, but are less fitting to assess how more narrow topics raise and drop in popularity. For tackling the latter problem, Fontana et al. (2019) and Önder et al. (2018) use a Latent Dirichlet Allocation (LDA) model, a part of the text analysis toolbox, to attribute papers to topics. However, LDA is notoriously unstable. For instance, referring to a typical presentation of LDA topics by their top words, Agrawal et al. (2018) remark that "due to LDA instability, the contents of such tables can only be described as mostly illusionary." Instead we perform our analysis using TF-IDF weights and a cosine similarity measure. That is an older and generally more noisy approach, but it comes with the benefit of stability and, hence, reproducibility.

Our findings can be contrasted with those by Fontana et al. (2019), who find that papers published in top-7 journals[2] have the average citation lag of 4.8–6.7 years, depending on the sample used for counting citations. Arguably, if topics formation is solely driven through citations, we would expect to find similar numbers using our methodology. The observation that our numbers for the durations between matched papers are substantially smaller, namely 2–4 years, suggests that common external factors also play a noticeable role in topic formation.

In summary, this paper shows that specific insights can be obtained by studying the temporal structure of the graph of textually similar papers, without the need for explicit topic modelling. Of course, explicit topic modelling can bring further insights but a typical procedure for topic modelling, namely LDA, is subject to instabilities. Possibly, state-of-the-art techniques such as the BERT model (Devlin et al., 2019), which are first pre-trained on a large general-purpose corpus and then fine-tuned by the researcher using the textual data of interest, could provide a more reliable construction of topics.

2 Data

We use the RePEc dataset, which is an open and distributed dataset with metadata on papers in economics.[3] Among other information, the RePEc dataset includes typical bibliographical information and abstracts for the majority of discussion and published papers. The results in this paper are based on the version of the dataset as downloaded on May 4, 2021. The data was downloaded using purpose-written code which is available on github.[4]

The dataset covers papers written in various languages. Besides English, some prominent languages include French, Russian, and German. While

[2]Traditional top-5 plus *International Economic Review* and *Review of Economics and Statistics*.

[3]`http://repec.org`

[4]`https://github.com/andrei-dubovik/repec`

there is allowance for data providers to indicate the language, this field is rarely filled in. Therefore, we have used an automatic language identification algorithm, namely Compact Language Detector 2.[5] For our analysis we only consider papers written in English.

Besides papers, the dataset also includes books, database records, and some other miscellaneous records. We filter on discussion papers and published papers. A substantial number of papers are non-economic papers or fringe economic papers. For instance, there are many purely medical papers in the dataset. We focus on economics papers and thus only select published and discussion papers that have JEL codes or that are published papers in the top 20 economic journals, following Combes and Linnemer (2010). Finally, we focus on years 1990–2020, and we exclude papers with missing titles or abstracts.

After the aforementioned filters have been applied, and after deduplication which we discuss in Appendix A, we end up with a total of 420,129 papers. Notably, there has been a sharp growth in the number of papers over the years—at least as found in the RePEc dataset—from 682 in 1990 to 29,875 in 2020 (with filters applied). We will have to account for this trend when doing analysis.

3 Analysis & Results

Methods of text analysis can be broadly divided into deterministic and stochastic methods. The former methods represent earlier approaches to text analysis, with one prominent example being the term frequency–inverse document frequency (TF-IDF, for short) weighting scheme combined with the cosine similarity measure. See Gentzkow et al. (2019) for an overview of the use of such deterministic methods in Economics literature.

Most of the contemporary computer science literature focuses on the stochastic methods, ranging from the popular word2vec model (Mikolov et al., 2013) to the state-of-the-art BERT model (Devlin et al., 2019). While earlier methods use a bag-of-words approach where a document is described by the frequencies of its constituent words, modern methods consider words together with their contexts. Modern approaches typically perform better on a broad range of natural language analysis tasks (see, e.g., Usherwood and Smit, 2019).

The better performance of modern methods comes at a cost. Modern stochastic methods are numerically complex. They typically feature between millions and thousands of millions parameters and highly non-linear objective functions. As a consequence, convergence is difficult to achieve. Instead, the recent computer science literature sidesteps this difficulty by simply applying an optimization procedure for a fixed number of steps in

[5] `https://github.com/CLD2Owners/cld2`

Table 1: Most Frequent Stems in Titles and Abstracts

Stem	Freq., %	Stem	Freq., %
paper	47.1	studi	30.9
use	46.5	find	30.5
result	36.4	market	27.4
model	34.5	show	27.0
effect	33.0	data	26.8

the direction of some local optima. In practice, this means that training those models with a different seed or slightly different data can produce markedly different results. Such variability in results effectively prevents the replicability of those results, an observation that some begin to emphasize in the computer science literature, see Wendlandt et al. (2018).

In contrast, deterministic methods are numerically stable and transparent as the only possible source of variance in the results comes from the outlined procedures. However, there is a considerable degree of freedom in the setup of deterministic methods and changing some of the aspects might produce different outcomes. After weighing the trade-offs, we have decided to refrain from using inherently unstable stochastic text analysis methods and instead employ a deterministic method, which we outline in detail below.

First, we encode each paper using a bag-of-words representation. From each paper we select its title and its abstract and concatenate them together, effectively giving equal weights to the words in the title and in the abstract. We then remove a limited number of stop-words (about 300 in total). Next, we select strings of consecutive Latin characters of length of at least 2 and treat these strings as words. In particular, all numbers that are encountered in titles or abstracts as well as JEL codes that are erroneously included in abstracts are excluded from the analysis. All words are then stemmed using the Porter's stemming algorithm (Porter, 1980). We remove any stems from the analysis that occur in 100 or fewer papers. All the remaining stems from all the papers are subsequently enumerated. Finally, we encode our data as matrix A, with dimensions 420,128×7,022, where A_{is} gives the number of occurrences of stem s in paper i.

A piece of trivia, namely a list with the most popular stems in the titles and abstracts of economics papers, is shown in Table 1. Also, an online tool for investigating single word trends as well as JEL code trends is available on the website of one of the authors.[6]

Let n denote the total number of papers. We apply the TF-IDF weight-

[6] `https://dubovik.eu/blog/repec`

ing scheme to matrix A to obtain a weighted matrix B:

$$B_{is} = \frac{\tilde{B}_{is}}{\|\tilde{B}_i\|_2}, \quad \tilde{B}_{is} = A_{is} \ln\left(\frac{n}{\sum_j \mathbb{I}(A_{js} > 0)}\right). \tag{1}$$

where $\mathbb{I}(A_{js} > 0)$ gives 1 if word s occurs in paper j at least once and 0 otherwise.

The TF-IDF weighting schemes gives higher weights to words that occur in fewer papers and are therefore likely to be more informative. As an example, consider the abstract of this paper. Matrix A weights every word by how often it occurs in the abstract, while excluding numbers, stop words, and those words that occur in 100 papers or less. If we use darker ink for the words with bigger weights in the matrix A, we obtain:[7]

---

study duration topics economics research looking much
time passes publication textually similar papers Using corpus
abstracts economics papers available dataset find
papers match papers year indicating strong common trends
economics literature Nevertheless matches long years apart
statistically significant suggesting topics last long Finally
average duration match dropped around years
years starting

Notably, words like "economics" and "paper" receive high weights because we use them often in our abstract. However, these words are quite common across all economic abstracts, therefore they receive lower weights under the TF-IDF weighting scheme. Instead, the latter emphasizes generally infrequent words such as "duration," "topic" or "corpus:"

---

study duration topics economics research looking much
time passes publication textually similar papers Using corpus
abstracts economics papers available dataset find
papers match papers year indicating strong common trends
economics literature Nevertheless matches long years apart
statistically significant suggesting topics last long Finally
average duration match dropped around years
years starting

For any two papers i and j we use the inner product (the cosine similarity) of the respective normalized TF-IDF vectors as a similarity measure,

$$m_{ij} = \langle B_i, B_j \rangle. \tag{2}$$

[7]As outlined, the analysis is done on the stems of words, but in the examples full word forms are preserved for readability.

Given that $\|B_i\| = 1$, we have $0 \leq m_{ij} \leq 1$. If m_{ij} is closer to 0, then few words between papers i and j are the same or those words that are the same have low weights, and so the papers are dissimilar. If m_{ij} is closer to 1, then papers i and j share the same words, and those words have high weights, and so the papers are similar.

A collection of papers with a distance function $d_{ij} = 1 - m_{ij}$ can be viewed as metric space. Alternatively, this collection of papers can be viewed as a complete weighted graph, with weights given by the similarity measure m. We adopt the latter view.

We reduce the complete graph by keeping only those edges that connect the closest papers. In doing so we effectively identify similar papers. There are two candidates for this procedure. Option A is that we define an absolute cutoff and then keep all those edges which connect papers with a similarity measure below the cutoff. Option B is that for each paper we select a given number of the nearest papers according to the similarity measure and then keep the respective edges. We refer to Option B as the relative cutoff.

For the comparison of the absolute and relative cutoffs, and for this comparison only, we use JEL codes. While it is not necessary that topics do not cross JEL code boundaries, it is to be expected that many topics fall within their own JEL codes. Therefore, among similar papers there should be many papers with the same JEL code (at least, if the procedure that identifies similar papers works). Consequently, we can use the precision coefficient—what share of similar papers share a JEL code—to compare the performance of various methods that identify similar papers. Formally, let $e_{ij} = 1$ if there is an edge between papers i and j, and let $e_{ij} = 0$ otherwise. Also, let $J_{ij} = 1$ if papers i and j have at least one JEL code in common, and let $J_{ij} = 0$ otherwise. We have

$$\text{Precision} = \frac{\sum_{ij} e_{ij} J_{ij}}{\sum_{ij} e_{ij}}. \tag{3}$$

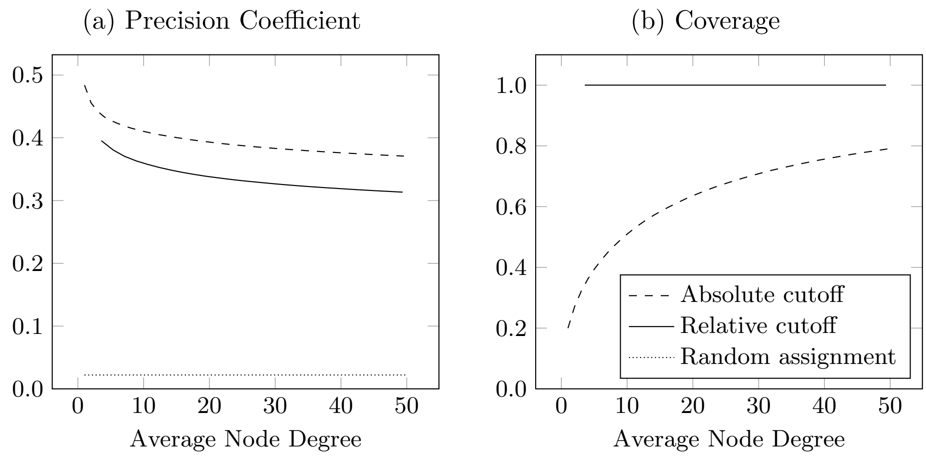
Fig. 1 shows the precision coefficient for the absolute cutoff, the relative cutoff, and for a random assignment of edges to papers (a benchmark scenario). On the x-axis is the average number of similar papers per any given paper, i.e. the average node degree. We observe that that both the absolute cutoff and the relative cutoff result in a substantially better precision than a random assignment. For instance, for an average node degree of 10 the absolute cutoff has precision at 41%, the relative cutoff has it at 36%, and the random assignment—at 2%.

Define coverage as the percentage of papers that have at least one similar paper, i.e.

$$\text{Coverage} = \frac{\sum_i \mathbb{I}(\sum_j e_{ij} > 0)}{n}. \tag{4}$$

While the absolute cutoff has a somewhat higher precision than the relative cutoff, it comes at the cost of a lower coverage. With an absolute cutoff

Figure 1: Comparison of Absolute and Relative Cutoffs



Notes: Precision coefficient is computed on the basis of a confusion matrix, where JEL codes serve as ground truth and TF-IDF matching with either relative or absolute cutoff is used for prediction. Coverage shows what percentage of papers have at least one similar paper when either relative or absolute cutoff is used. Average node degree is the average number of similar papers for any given paper.

it might in principle happen that some papers have no neighbours with a similarity measure below the cutoff. In fact, it happens quite often as Fig. 1, right panel, demonstrates. For instance, for an average node degree of 10 the relative cutoff results in a 100% coverage by definition, while the absolute cutoff has it at only 51%. If we adopt a procedure with lower coverage, we effectively have fewer observations when we do the upcoming analysis, which decreases efficiency and potentially introduces a selection bias. For this reason we adopt the relative cutoff instead of the absolute cutoff. As we focus on small topics, we set the relative cutoff at the 5^{th} closest paper. As a robustness check, we also run the complete analysis when the cutoff is set at 1 or 10 nearest neighbours.

Formally, we connect papers i and j if and only if paper j is among the 5 closest neighbours of paper i or paper i is among the 5 closest neighbours of paper j. Let e_{ij} denote a weighted edge between papers i and j. Also, let $d_{i(k)}$ denote the k^{th} order statistic of vector $(d_{ij})_j$, and analogously for $d_{(k)j}$. We set:

$$e_{ij} = \begin{cases} m_{ij} & \text{if} \quad \big(d_{ij} \leq d_{i(5)}\big) \vee \big(d_{ij} \leq d_{(5)j}\big), \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

This defines a weighted graph $\Gamma = (V, \{ij | e_{ij} > 0\})$ with weights e_{ij}, where V is the set of all papers. The reduced graph Γ is smaller than the complete graph: instead of $\approx 8.8 \cdot 10^{10}$ edges in the complete graph we obtain

$\approx 1.9 \cdot 10^6$ edges in the reduced graph, a number that is easier to work with from a technical perspective.

An example of a neighbourhood from graph Γ is shown in Fig. 2. Nodes are papers and edges connect similar papers. The selected neighbourhood is around (a) Angrist et al. (2017), "Economic Research Evolves: Fields and Styles." The distances in the figure were chosen to correlate in rank with d_{ij}, but a perfect embedding in R^2 is not possible, so the rank correlation is less than one. We find that many connected papers conduct citation analysis, with several of those putting an emphasis on within-field and cross-field outcomes. However, we find erroneous matches as well, e.g. (o) Davis and Weinstein (2001) is a review of key empirical findings in international trade. Still, as long as such mistakes are random and do not depend on publication years, our findings regarding the temporal relations between papers will be unbiased. The figure further illustrates that we draw an edge between two papers if either of them is among the top 5 neighbours of another. For instance, while (n) is not among the top 5 closest neighbours of (a), (a) is among the top 5 closest neighbours of (n), and so we draw an edge.

If $e_{ij} > 0$ and $e_{jk} > 0$ but $e_{ik} = 0$, we do not consider papers i and k as connected. That is, if i is textually similar to j and j is textually similar to k, we nevertheless do not connect papers i and k if they are not textually similar to one another to start with. Alternatively, we can consider any two papers connected if there is a simple path between the two papers of at most a given length p. Formally, we can consider a weighted graph $\Gamma^{(p)}$ with edges $e_{ik}^{(p)}$ defined recursively as

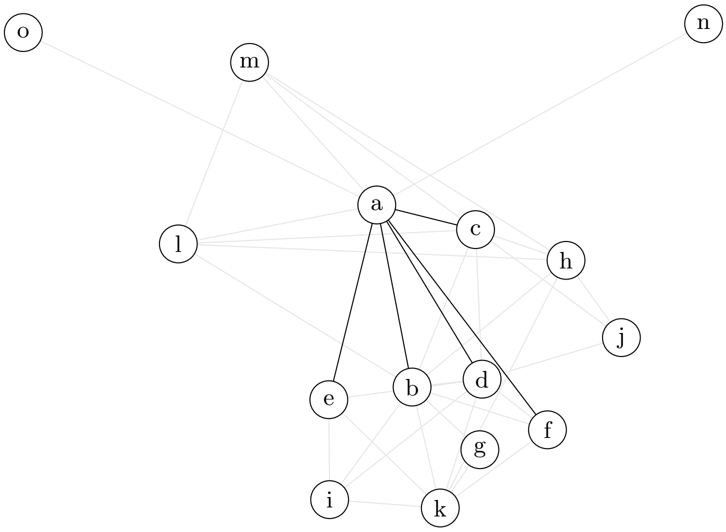
$$e_{ik}^{(p)} = \max_{k,\, 1 \le q \le p} \min \left(e_{ij}^{(q)}, e_{jk}^{(p-q)} \right), \tag{6}$$

where $e_{ij}^{(1)} = e_{ij}$ and $e_{ij}^{(0)} = 1$. In Appendix B we consider paths of length 2 and 3 and show that our findings are robust to this extended specification of what papers are considered connected.

Throughout this paper the term "topic" is understood to be a connected subgraph of graph Γ. All our results can be derived without an explicit construction of topics. We therefore forgo doing so, as explicit topic construction requires imposing relatively strict assumptions on the data generating process for the papers and topics, and the results could potentially be sensitive to those assumptions. Also, given that the construction of Γ does not depend on JEL codes, all the derived results are independent of JEL codes as well.

Let us consider the temporal relations between similar papers. Let $y(i)$ denote the year in which paper i was first released or published, and let ε_u denote a column vector with 1 at position u and zeroes elsewhere. We

Figure 2: An Example of a Neighbourhood in Γ



Notes: The figure shows top 10 nearest neighbours of Angrist et al. 2017—node (a)—as well as those papers who have Angrist et al. 2017 among their top 5 nearest neighbours. If there is an edge between two papers, it is shown as a gray line; if the edge is between Angrist et al. 2017 and one of its top 5 nearest neighbours, it is shown in black. The Euclidian distances in the figure rank-correlate with d_{ij}. The papers are as follows: (a) Angrist et al. 2017, (b) Ketzler and Zimmermann 2013, (c) Adams et al. 2004, (d) Galiani and Gálvez 2017, (e) Anauati et al. 2020, (f) Nedelchev 2017, (g) McCabe and Mueller-Langer 2019, (h) Kim et al. 2011, (i) Meyer et al. 2018, (j) Finardi 2017, (k) Hamermesh 2018, (l) Angrist et al. 2020, (m) Angrist et al. 2020, (n) Ioan-Franc 2003, (o) Davis and Weinstein 2001. Nodes (l) and (m) reference two different versions of the same paper, because the authours are spelled substantially differently and so this case was not caught by the deduplication procedure: one version uses "Josh Angrist" and another uses "Joshua Angrist".

construct a symmetric matrix

$$P = \sum_{i,j} \varepsilon_{y(i)} \varepsilon'_{y(j)} \, e_{ij} \Big/ \sum_{i,j} e_{ij}. \tag{7}$$

P has dimensions 31×31. If we pick two similar papers at random, say paper i and paper j, then P_{uv} gives the (weighted) probability that paper i is from year u and paper j is from year v (or the other way around, that paper i is from year v and paper j is from year u).

As has been mentioned earlier, most papers in the RePEc database come from later years. Consequently, most probability mass in matrix P is also concentrated in later years. We can negate this bias with an appropriately chosen normalization. Namely, consider matrix $Q = P\mathbf{J}P'$, where $\mathbf{J}$ denotes a matrix of ones. Q_{uv} gives the (weighted) probability that in an ordered pair of randomly chosen papers the first paper would be from year u and the second—from year v.

We then normalize P with Q, element-wise. We also apply the log transformation so as to increase the differential entropy of the resulting numbers. This transformation improves contrast in Figures 3 and 5 and, potentially, the power of the statistical tests. So, we compute

$$L = \log(P) - \log(Q). \tag{8}$$

The resulting matrix L is shown in Fig. 3, left panel. Before discussing the figure, we perform a placebo test to ensure that the results are not accidentally obtained by construction (e.g., due to an oversight in the method or a bug in the code).

Let σ_k denote a random permutation of the papers indexed by k. We use this permutation to define function $\tilde{y}_k(i)$ that assigns paper i a random year from the set of all the years of all the papers, in effect reshuffling the years across papers:

$$\tilde{y}_k(i) = \sigma_k\Big((y(1), \ldots, y(n))\Big)_i. \tag{9}$$

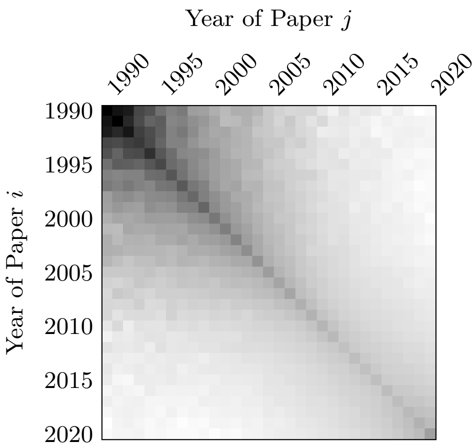
We proceed to construct matrix $\tilde{L}_k$ in exactly the same way as we have constructed L but replacing $y(\cdot)$ with $\tilde{y}_k(\cdot)$. The resulting matrix, for some draw k is shown in Fig. 3, right panel.

Besides serving as a placebo test for the code, the process of generating $\tilde{L}_k$ can be viewed as a data generating process for L under the null hypothesis that the topic on which any paper gets written is independent of the year of that paper.

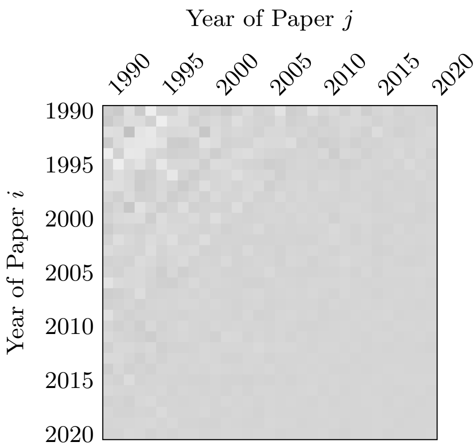
We observe that under the placebo test there are no patterns in the figure, as expected. (The stronger noise in the upper left corner corresponds to the few papers being available in those years.) However, in the actual data we see a pronounced clustering along the diagonal. Each year papers are written predominantly on the topic or topics popular in that particular year. While this result would likely not come as surprising to a practicing economist—it is easy to speculate on possible explanations—it is a novel result to the best of our knowledge and, equally important, it need not have been observed. We might have observed instead a figure where both the diagonal and the off-diagonals are pronounced meaning that most topics stay popular for several years. We see some of that possibility in the earlier years—compare the upper left to the lower right corner—and we discuss it in more detail later on.

We proceed with testing which topic durations are consistent with the data. To do so, we keep the same null hypothesis $\mathcal{H}_0$ that topics are independent of years but we test it against different alternative hypotheses. Namely, consider hypothesis $\mathcal{H}_a^t$ that there are some topics that stay popular for exactly t years. We allow for $t = 0$ to refer to topics that are popular for exactly 1 *calendar* year. (If papers are uniformly distributed within a year,

Figure 3: Topic Incidences (All Papers)



(a) Original Data



(b) Bootstrapped Data

Notes: Each panel shows adjusted frequencies of edges between papers, tabulated by the years of those papers. Panel (a) shows the actual frequencies, whereas Panel (b) shows simulated frequencies when the years of papers are randomly perturbed.

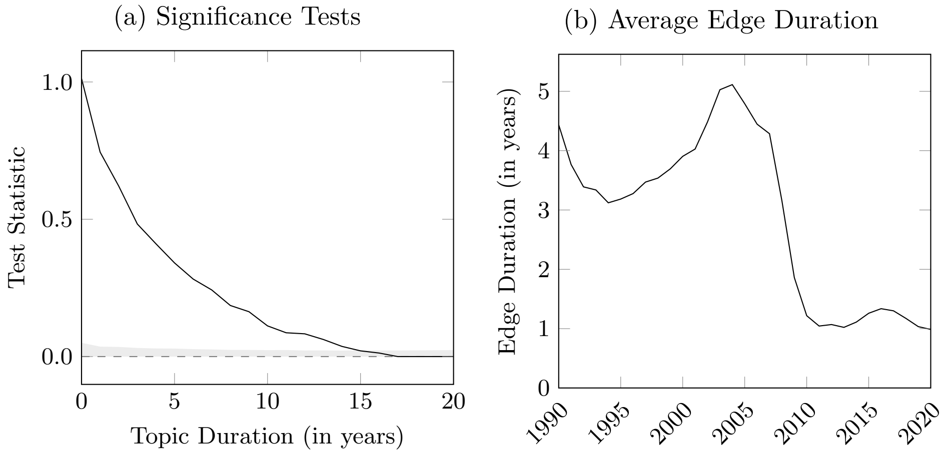
then the average distance between them is $1/3 \approx 0$ years.) Under $\mathcal{H}_a^t$, some of the papers that are t years apart or less should match more frequently than random, while that should not be the case for papers that are more than t years apart. I.e., it should be the case that $L_{uv} > 0$ for some u, v such that $|u - v| \leq t$ and it should be that $L_{uv} \leq 0$ whenever $|u - v| > t$. There are a number of ways to construct a test statistic that captures these considerations. We simply test for the presence of some positive L_{uv} on the boundary where $|u - v| = t$. To that end we use the following test statistic,

$$x^t = \sum_{u=1}^{T-t} \max(L_{u(u+t)}, 0). \tag{10}$$

We compute the distribution of x^t under $\mathcal{H}_0$ using bootstrapping. We generate 1,000 random permutations σ_k and we compute $\tilde{x}_k^t$ analogously with x^t but using $\tilde{L}_k$ instead of L. Then the 95$^{\text{th}}$ percentile of $\left\{\tilde{x}_k^t\right\}_k$ is used as a one-sided critical value. We do not correct for multiple comparisons, because the alternative hypotheses are approximately nested: if there are topics popular for t years, we are nearly as likely to see matches between papers that are $t-1$ years apart as if there only were topics that are popular for $t - 1$ years. The results are shown in Fig. 4, left panel, as a graph of x_a^t over t, together with the corresponding critical values.

We see that the test statistic is larger for shorter durations. However, statistically we can reject the null that topics are independent of publication

Figure 4: Topic Duration (All Papers)



Notes: the left panel shows a test statistic for an $\mathcal{H}_0$ that topics are independent of publication year against an alternative $\mathcal{H}_a^t$ that some topics stay popular for exactly t years, where t is shown on the horizontal axis. The grey area shows the confidence interval where $\mathcal{H}_0$ cannot be rejected at 5%. The right panel gives the average weighted duration between two neighbouring papers, where at least one of the papers has been published in the given year.

years in favour of an alternative hypothesis that there are some topics that last as long as 14 years.

How did the average topic duration change over time? We cannot answer this question exactly without an explicit clustering of graph Γ into topics. That said, under the assumption that the expected edge duration for any edge from a given topic is indicative of that topic's duration, we can instead look at how edge duration changed over time.

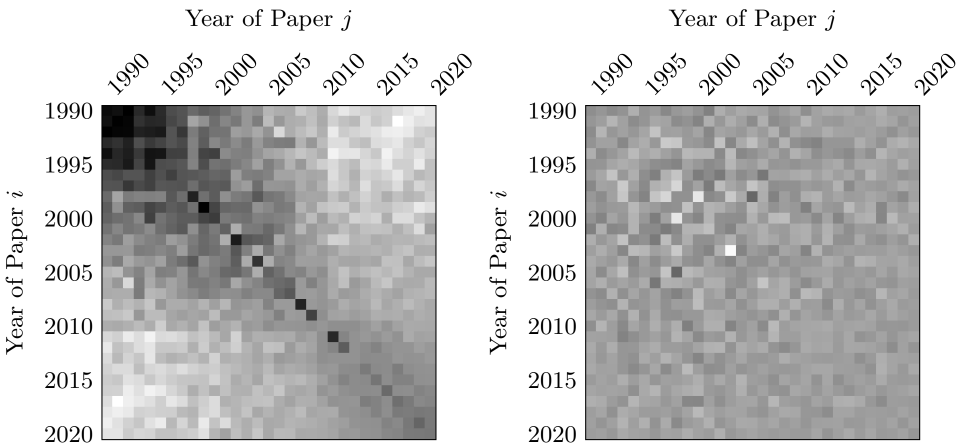
To stay consistent with the computation of x^t, we will use the positive elements of L as weights. For instance, if between year u and year τ there were twice as many matches as would otherwise happen randomly, we assign a weight of $\log(2)$ to the duration $|u - \tau|$. And if there were fewer matches than what would happen randomly, we assign a weight of 0. So, we compute a weighted average edge duration as follows:

$$t_\tau = \frac{\sum_{u=1}^{T} \max(L_{u\tau}, 0) \cdot |u - \tau|}{\sum_{u=1}^{T} \max(L_{u\tau}, 0)}. \tag{11}$$

The results are shown in Fig. 4, right panel. Notably, the average edge duration has dropped from somewhere around 3–5 years in 1990–2005 to about 1 year beginning 2010. In absolute terms, edge duration need not equal topic duration, but the relative drop in the average edge duration can still suggest that the same drop has occurred for the average topic duration.

A gradual decrease in the topic duration would have been consistent

Figure 5: Topic Incidences (Top 20)



(a) Original Data (b) Bootstrapped Data

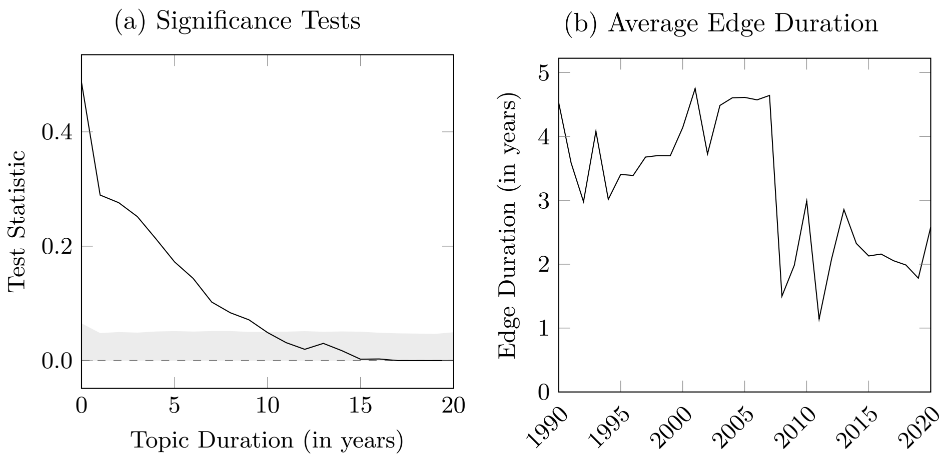
Notes: Each panel shows adjusted frequencies of edges between papers, tabulated by the years of those papers. Panel (a) shows the actual frequencies, whereas Panel (b) shows simulated frequencies when the years of papers are randomly perturbed.

with a gradual increase in the pace of the economics research. Contrary, what we observe in Fig. 4, is a relatively rapid shift in the topic duration. One possible explanation for such a shift is the drop in the search costs for related working papers following the launch of Google Scholar at the end of 2004. As easier discovery of related working papers likely leads to the positioning of any given paper more in line with the most recent research, we would indeed observe shorter topic duration. Having said this, absent usage statistics of Google Scholar we cannot test this logic further and so we leave it as a conjecture.

The results so far have been based on discussion papers, of which about 30,000 are written yearly as of late. For comparison, we repeat the whole exercise but limit the corpus to the top tier research. For that purpose, we use papers published in the top 20 journals according to Combes and Linnemer (2010). While any journal ranking beyond the Top 5 is open for debate, we have chosen to adopt a specific ranking so as to increase the sample size to more journals than just five. Based on this selection we obtain matrix L as given in Fig. 5, left panel. The average edge duration given in Fig. 6, right panel.

We observe that the longest statistically significant topic duration is now shorter, but that result is due to substantially fewer papers and the consequent loss of efficiency of our estimators. The average edge duration have dropped noticeably less than it has done for all papers in economics, from about 3–4 year in 1990–2005 to about 2 years as of late. It must

Figure 6: Topic Duration (Top 20)



Notes: the left panel shows a test statistic for an $\mathcal{H}_0$ that topics are independent of publication year against an alternative $\mathcal{H}_a^t$ that some topics stay popular for exactly t years, where t is shown on the horizontal axis. The grey area shows the confidence interval where $\mathcal{H}_0$ cannot be rejected at 5%. The right panel gives the average weighted duration between two neighbouring papers, where at least one of the papers has been published in the given year.

be mentioned, however, that when analysing all papers, the majority of papers are discussion papers, in which case $y(\cdot)$ gives the year when those discussion papers were released. On the other hand, here we look exclusively at published papers and $y(\cdot)$ gives the publication year, which can have a higher spread due to the variance in publication lag. As such, a 2 year average edge duration in Fig. 6 is not per se inconsistent with a 1 year edge duration from Fig. 4. So, seemingly, the average edge duration as well as, potentially, the average topic duration have declined overall, whether it's top-tier research papers or economic research at large.

As a robustness check, we repeat the whole analysis when the cutoff in Eq. (5) is set at 1 or 10 nearest neighbours, see Appendix B. We obtain qualitatively identical results.

4 Summary

In conclusion, we briefly recap the main points of our analysis. We have started by collecting and cleaning the RePEc dataset with the metadata on economics papers. The corresponding code has been made publicly available.[8] There is also an online tool for visualizing how often a specific word (or two words) have been mentioned over time in the RePEc papers.[9] Ad-

[8] `https://github.com/andrei-dubovik/repec`
[9] `https://dubovik.eu/blog/repec`

ditionally to cleaning the data, we have also identified duplicates, which would typically be a discussion paper and its published version. When a paper has multiple versions, we always use the earliest available one in our analysis, and that insulates the analysis from the effects of the publication lag. Building on the TF-IDF metric, we have constructed a measure of text similarity for every pair of papers. The measure is based on papers' titles and abstracts and, importantly, we have explicitly removed any possible mention of years or JEL codes from those texts.

Given our measure of text similarity, we tabulate how many similar papers there are between different pairs of years. We find strong evidence that most similar papers are published in the same year, which says there are strong common trends in the economics literature. We additionally observe that the average difference in publication years between two similar papers has dropped from some 4 years during 1990–2005 to about 1 year starting in 2010. We conjecture that a possible explanation is the launch of Google Scholar at the end of 2004.

Specifically for our case, we have also developed a statistical test that allows us to asses whether the observed frequency of similar papers published a given number of years apart is statistically significant. We bootstrap the null statistic for our test by perturbing the assignment of publication years to papers. While the observed higher frequency of similar papers from the same year is indeed statistically significant, we also find that matches as long as 14 years apart are statistically significant as well. This latter result suggests there are topics in the economics literature that last at least as long.

Finally, we conduct a number of robustness checks: we check for a subsample of papers that are published in the top 20 journals, we vary a cutoff for our similarity measure so that fewer or more papers are found similar, we extend our similarity measure to include papers that are indirectly similar to one another through a chain of other papers. In all of these robustness checks we find qualitatively very similar results.

References

Adams, J. D., Clemmons, J. R., and Stephan, P. E. (2004). Standing on academic shoulders: Measuring scientific influence in universities. NBER Working Paper 10875.

Agrawal, A., Fu, W., and Menzies, T. (2018). What is wrong with topic modeling? And how to fix it using search-based software engineering. *Information and Software Technology*, 98:74–88.

Anauati, M. V., Galiani, S., and Gálvez, R. H. (2020). Differences in citation patterns across journal tiers: The case of economics. *Economic Inquiry*, 58:1217–1232.

Angrist, J., Azoulay, P., Ellison, G., Hill, R., and Lu, S. F. (2017). Economic research evolves: Fields and styles. *American Economic Review*, 107(5):293–97.

Angrist, J., Azoulay, P., Ellison, G., Hill, R., and Lu, S. F. (2020). Inside job or deep impact? Extramural citations and the influence of economic scholarship. *Journal of Economic Literature*, 58(1):3–52.

Bellas, A. and Kosnik, L.-R. (2019). Which leading journal leads? Idea diffusion in economics research journals. *Empirical Economics*, 57(3):901–921.

Card, D. and DellaVigna, S. (2013). Nine facts about top journals in economics. *Journal of Economic Literature*, 51(1):144–61.

Cherrier, B. (2017). Classifying economics: A history of the jel codes. *Journal of economic literature*, 55(2):545–79.

Combes, P.-P. and Linnemer, L. (2010). Inferring missing citations: A quantitative multi-criteria ranking of all journals in economics. GREQAM discussion paper 2010-28.

Davis, D. R. and Weinstein, D. E. (2001). What role for empirics in international trade? NBER Working Paper 8543.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 4171–4186.

Finardi, U. (2017). Long time series of highly cited articles: an empirical study. IRCrES Working Paper 12/2017.

Fontana, M., Montobbio, F., and Racca, P. (2019). Topics and geographical diffusion of knowledge in top economic journals. *Economic Inquiry*, 57(4):1771–1797.

Galiani, S. and Gálvez, R. H. (2017). The life cycle of scholarly articles across fields of research. NBER Working Paper 23447.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–574.

Hamermesh, D. S. (2013). Six decades of top economics publishing: Who and how? *Journal of Economic Literature*, 51(1):162–72.

Hamermesh, D. S. (2018). Citations in economics: Measurement, uses, and impacts. *Journal of Economic Literature*, 56:115–156.

Ioan-Franc, V. (2003). The evaluation of the economic research works. A relevance indicator. *Romanian Journal of Economic Forecasting*, 4:37–48.

Ketzler, R. and Zimmermann, K. F. (2013). A citation-analysis of economic research institutes. *Scientometrics*, 95:1095–1112.

Kim, J.-Y., Min, I., and Zimmermann, C. (2011). The economics of citation. *The Korean Economic Review*, 27:93–114.

McCabe, M. J. and Mueller-Langer, F. (2019). Does data disclosure increase citations? Empirical evidence from a natural experiment in leading economics journals. Available online at SSRN.

Meyer, M., Waldkirch, R. W., Duscher, I., and Just, A. (2018). Drivers of citations: An analysis of publications in "top" accounting journals. *Critical Perspectives on Accounting*, 51:24–46.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Nedelchev, M. (2017). A bibliometric study of citations in corporate governance. *Entrepreneurship*, V:95–105.

Önder, A. S., Popov, S. V., and Schweitzer, S. (2018). Leadership in scholarship: A machine learning based investigation of editors' influence on textual structure. Technical report, Cardiff Economics Working Papers.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Usherwood, P. and Smit, S. (2019). Low-shot classification: A comparison of classical and deep transfer machine learning approaches. arXiv preprint arXiv:1907.07543.

Wendlandt, L., Kummerfeld, J. K., and Mihalcea, R. (2018). Factors influencing the surprising instability of word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 2092–2102.

Appendix A: Deduplication

The RePEc dataset frequently contains multiple iterations of the same paper. For instance, a paper could be released one or more time as a working paper and then once as a published paper. To avoid double counting and possible spurious results we employ a two-stage deduplication procedure. In a nutshell, we identify duplicate papers based on authors' names, titles, and abstracts.

The first step consists of pairing papers that share the same authors. Doing so is not trivial, because for many authors the spelling of their names differs between papers. For example, initials may be abbreviated, middle-names omitted, and the order of the name parts changed. Furthermore, the names come from different languages and cultures making it impractical to distinguish surnames, given names, middle names, and prepositions. Instead, we employ a method that can match different spellings of the same name without relying on the semantic understanding of that name.

Consider an arbitrary name. First, we clean the name by removing most non-Latin characters and harmonizing the spelling of hyphens and apostrophes.[10] Second, we split the author's name into a set of lower case words (tokens). Third, we generate a list of possible derived names for the author. A derived name is a name that can be obtained by applying the following two steps: (1) abbreviating up to all but one token to a single letter, (2) omitting up to all but two tokens, one of which needs to be a non-abbreviated token. By definition, the set of derived names includes the name itself. If a name contains 8 or more tokens we limit the set of derived tokens to the name itself.

For example, "Keynes, John Maynard" is first translated to a set of three tokens: {john, keynes, maynard}. From these 3 tokens 14 derived names are generated, e.g., {j, k, maynard}, {john, keynes, m}, {j, maynard}, etc.

Two authors match if one of them can be derived from the other. Formally, let $\mathrm{ch}(r)$ be the set of all derived names for author r. Let $\mathrm{ma}(r, s)$ indicate if authors r and s match. We have

$$\mathrm{ma}(r, s) = \begin{cases} 1 & \text{if } r \in \mathrm{ch}(s) \vee s \in \mathrm{ch}(r), \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

Consider two papers, i and j. We say these papers match by author if and only if (1) both papers have the same number of authors and (2) there is a one-to-one match between the authors' names. Formally, let mp_{ij} indicate whether papers i and j match by author, let a^i denote a tuple of the authors of paper i so that a^i_k is the k-th author, and let σ denote a permutation.

[10] As we only consider papers written in English we feel that limiting the names to Latin characters (including diacritics) is appropriate.

We have

$$\mathrm{mp}_{ij} = \begin{cases} 1 & \text{if } |a^i| = |a^j| \;\wedge\; \exists \sigma : \mathrm{ma}(\sigma(a^i)_k, a^j_k) = 1 \;\forall k, \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

Additionally to matching by author, we match papers by title and abstract. Towards this end we construct cosine similarity measure m_{ij} between papers i and j that is computed on the basis of a TF-IDF representation of the word stems from titles and abstracts. We discuss this procedure in detail in Section 3. We say that papers i and j match by title and abstract if either their titles are exactly the same or $m_{ij} > 0.75$.

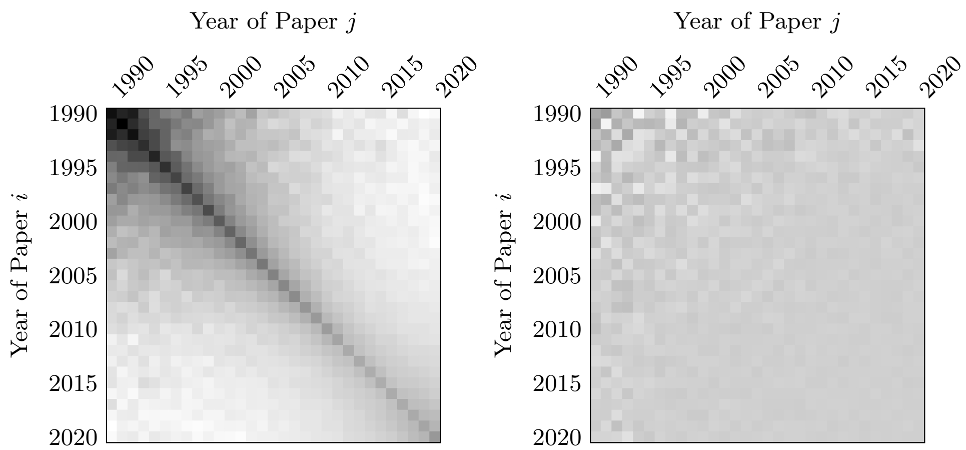
Finally, we combine both measures. We say that two papers match if they match both by author and by title and abstract.

The threshold of 0.75 is based on a manual inspection of the results and the results are robust towards varying it. Choosing a threshold that is too large leads to too few matches and to the presence of duplicate papers in the final dataset. Given that iterations of a paper are rarely released in the same year, the presence of duplicates increases the persistence of topics across years. In contrast, a threshold that is too low may lead to spurious matches for some papers. That being said, combining author information and content information mitigates the risk of spurious matches.

The procedure outlined so far gives us pairs of papers that have been identified as duplicates. In general, papers can undergo multiple revisions where the changes between each revision are small while the changes from the first revision to the last could be large. To identify chains of revisions it is useful to interpret the identified duplicate pairs as edges in an graph. Based on this graph we consider each connected component a unique paper such that if we identified i, j and j, k as matches we consider i, j, k as versions of the same paper.

Appendix B

Figure 7: Topic Incidences (All Papers, 1 Nearest Neighbour)

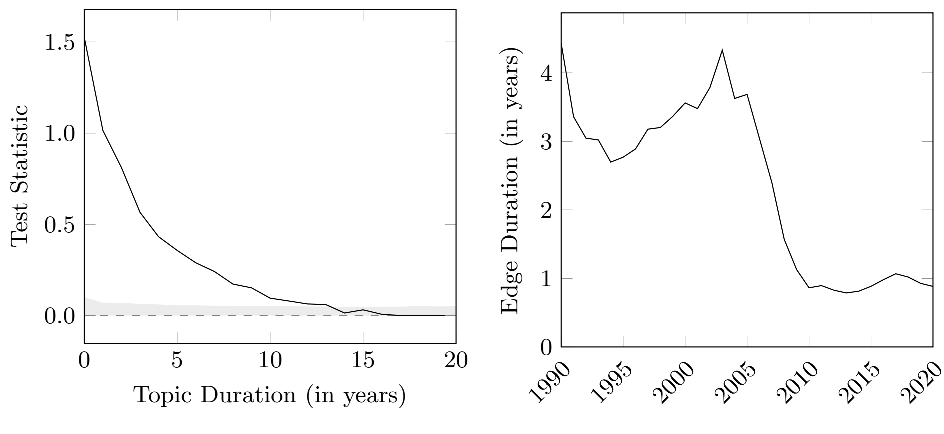


(a) Original Data (b) Bootstrapped Data

Notes: Each panel shows adjusted frequencies of edges between papers, tabulated by the years of those papers. Panel (a) shows the actual frequencies, whereas Panel (b) shows simulated frequencies when the years of papers are randomly perturbed.

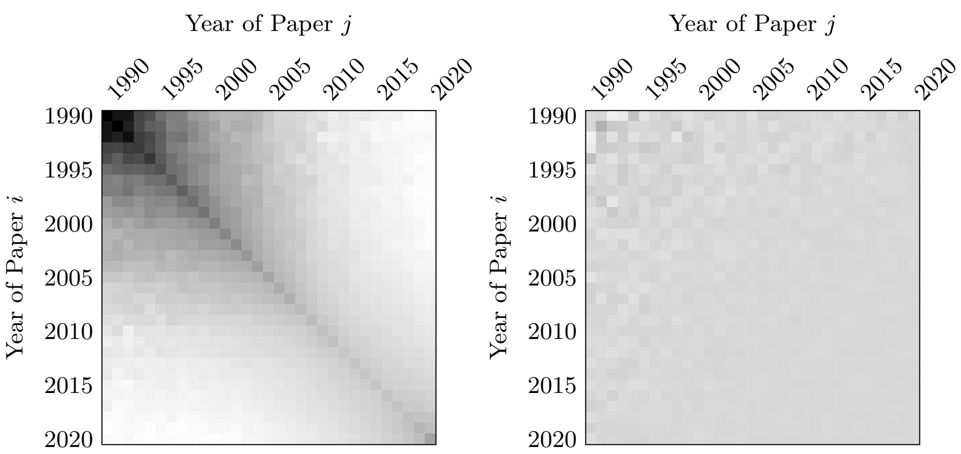
Figure 8: Topic Duration (All Papers, 1 Nearest Neighbour)

(a) Significance Tests (b) Average Edge Duration



Notes: the left panel shows a test statistic for an $\mathcal{H}_0$ that topics are independent of publication year against an alternative $\mathcal{H}_a^t$ that some topics stay popular for exactly t years, where t is shown on the horizontal axis. The grey area shows the confidence interval where $\mathcal{H}_0$ cannot be rejected at 5%. The right panel gives the average weighted duration between two neighbouring papers, where at least one of the papers has been published in the given year.

Figure 9: Topic Incidences (All Papers, 10 Nearest Neighbours)

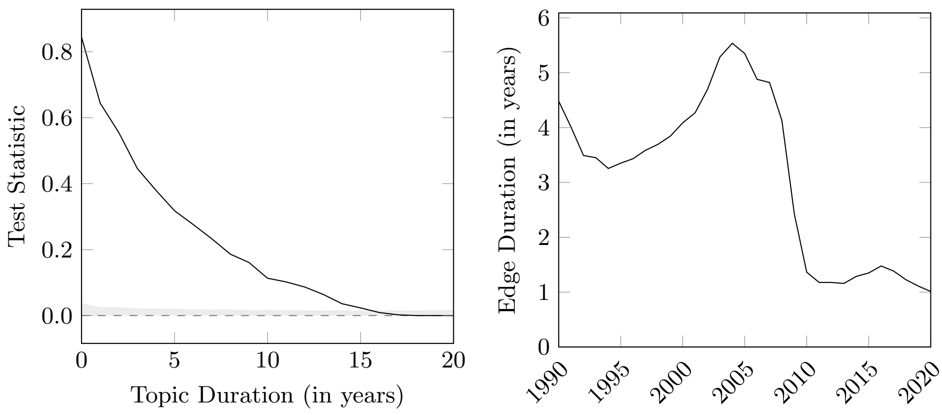


(a) Original Data (b) Bootstrapped Data

Notes: Each panel shows adjusted frequencies of edges between papers, tabulated by the years of those papers. Panel (a) shows the actual frequencies, whereas Panel (b) shows simulated frequencies when the years of papers are randomly perturbed.

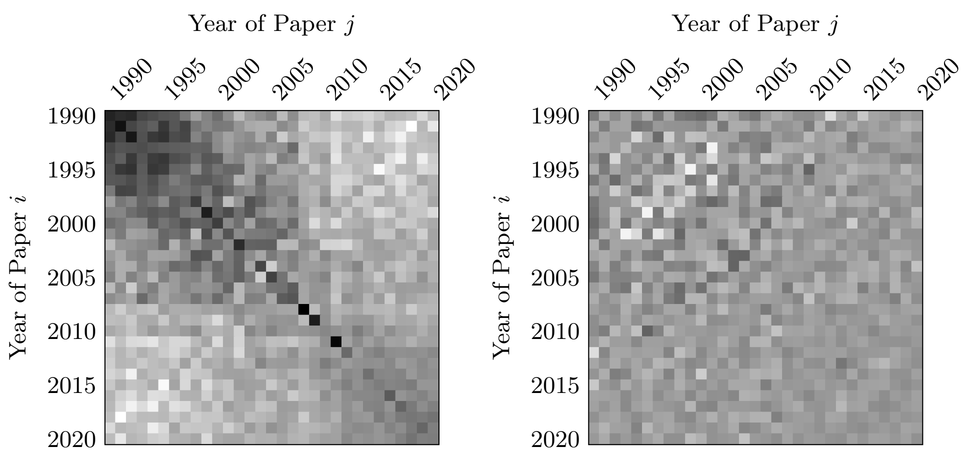
Figure 10: Topic Duration (All Papers, 10 Nearest Neighbours)

(a) Significance Tests (b) Average Edge Duration



Notes: the left panel shows a test statistic for an $\mathcal{H}_0$ that topics are independent of publication year against an alternative $\mathcal{H}_a^t$ that some topics stay popular for exactly t years, where t is shown on the horizontal axis. The grey area shows the confidence interval where $\mathcal{H}_0$ cannot be rejected at 5%. The right panel gives the average weighted duration between two neighbouring papers, where at least one of the papers has been published in the given year.

Figure 11: Topic Incidences (Top 20, 1 Nearest Neighbour)



(a) Original Data

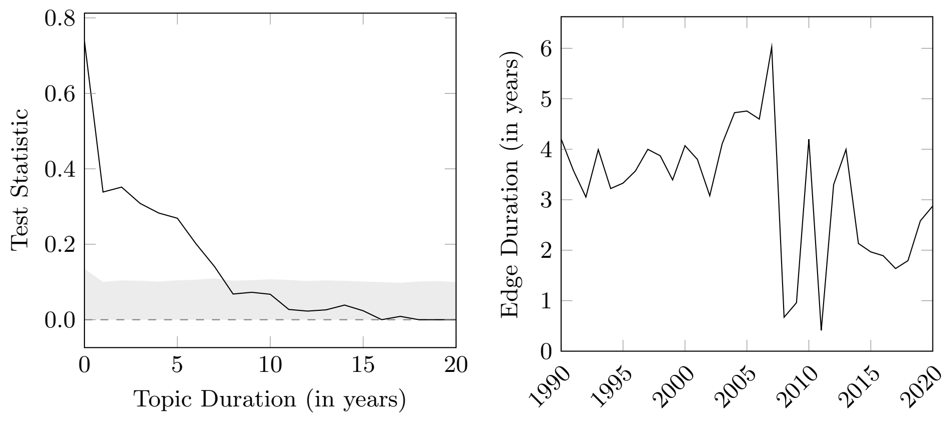
(b) Bootstrapped Data

Notes: Each panel shows adjusted frequencies of edges between papers, tabulated by the years of those papers. Panel (a) shows the actual frequencies, whereas Panel (b) shows simulated frequencies when the years of papers are randomly perturbed.

Figure 12: Topic Duration (Top 20, 1 Nearest Neighbour)

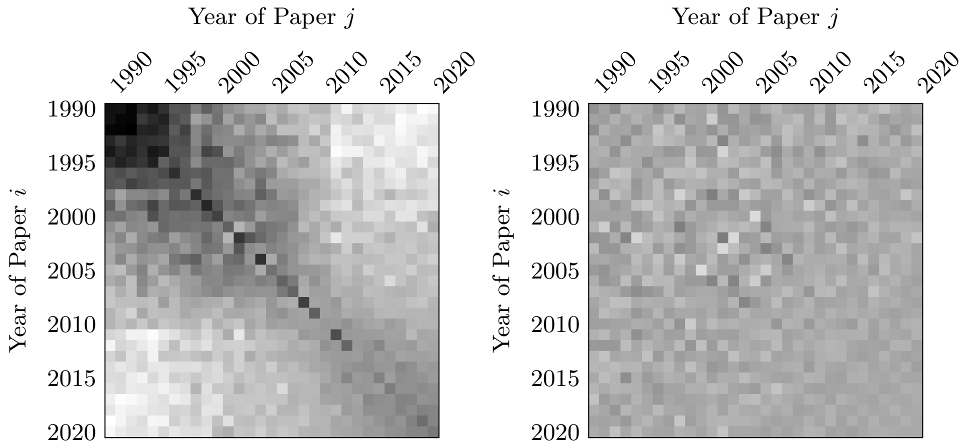
(a) Significance Tests

(b) Average Edge Duration



Notes: the left panel shows a test statistic for an $\mathcal{H}_0$ that topics are independent of publication year against an alternative $\mathcal{H}_a^t$ that some topics stay popular for exactly t years, where t is shown on the horizontal axis. The grey area shows the confidence interval where $\mathcal{H}_0$ cannot be rejected at 5%. The right panel gives the average weighted duration between two neighbouring papers, where at least one of the papers has been published in the given year.

Figure 13: Topic Incidences (Top 20, 10 Nearest Neighbours)

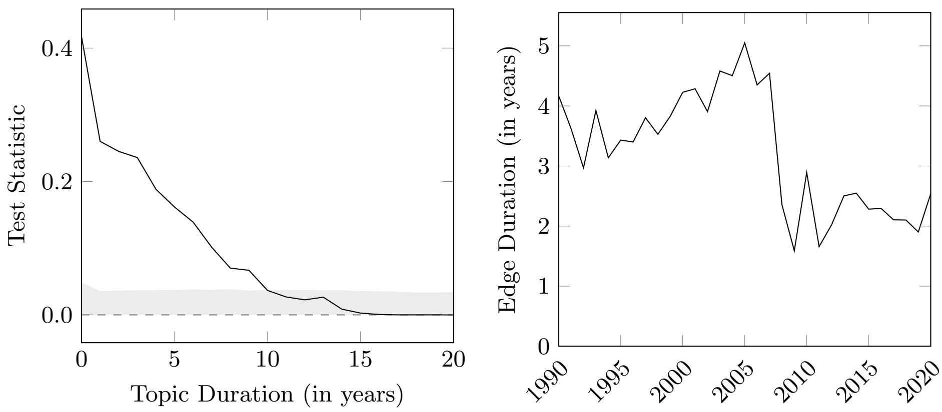


(a) Original Data (b) Bootstrapped Data

Notes: Each panel shows adjusted frequencies of edges between papers, tabulated by the years of those papers. Panel (a) shows the actual frequencies, whereas Panel (b) shows simulated frequencies when the years of papers are randomly perturbed.

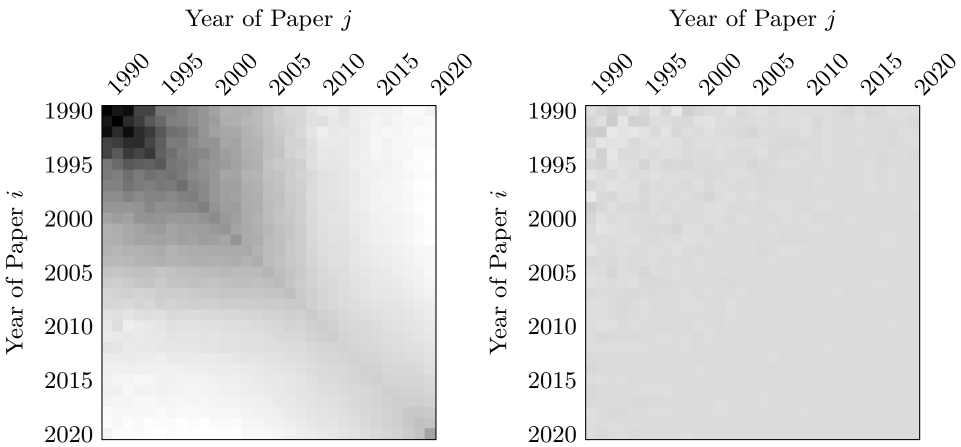
Figure 14: Topic Duration (Top 20, 10 Nearest Neighbours)

(a) Significance Tests (b) Average Edge Duration



Notes: the left panel shows a test statistic for an $\mathcal{H}_0$ that topics are independent of publication year against an alternative $\mathcal{H}_a^t$ that some topics stay popular for exactly t years, where t is shown on the horizontal axis. The grey area shows the confidence interval where $\mathcal{H}_0$ cannot be rejected at 5%. The right panel gives the average weighted duration between two neighbouring papers, where at least one of the papers has been published in the given year.

Figure 15: Topic Incidences (All Papers, Paths of Length 2)



(a) Original Data

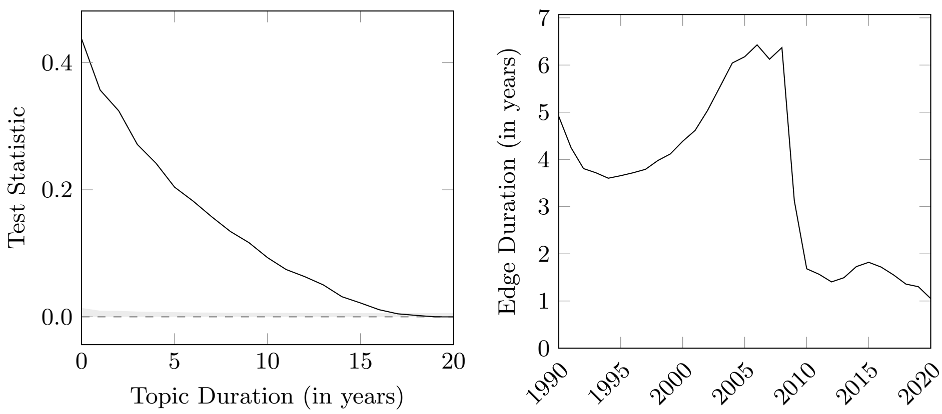
(b) Bootstrapped Data

Notes: Each panel shows adjusted frequencies of edges between papers, tabulated by the years of those papers. Panel (a) shows the actual frequencies, whereas Panel (b) shows simulated frequencies when the years of papers are randomly perturbed.

Figure 16: Topic Duration (All Papers, Paths of Length 2)

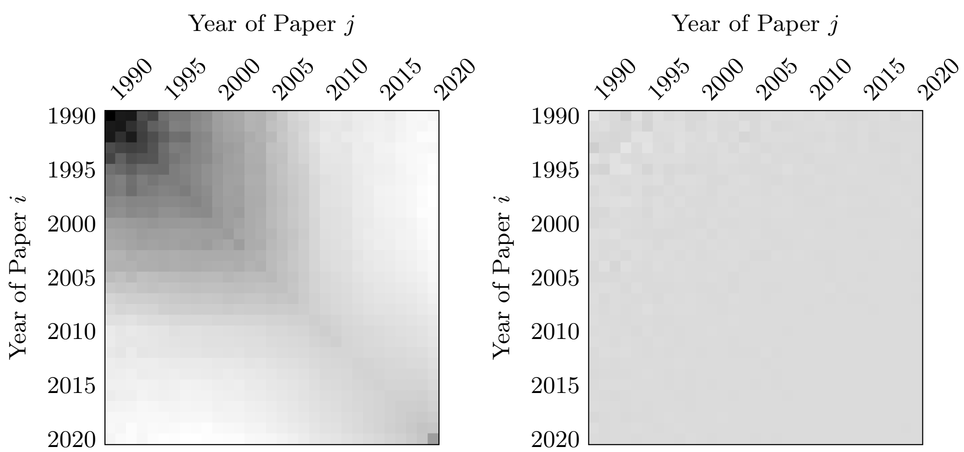
(a) Significance Tests

(b) Average Edge Duration



Notes: the left panel shows a test statistic for an $\mathcal{H}_0$ that topics are independent of publication year against an alternative $\mathcal{H}_a^t$ that some topics stay popular for exactly t years, where t is shown on the horizontal axis. The grey area shows the confidence interval where $\mathcal{H}_0$ cannot be rejected at 5%. The right panel gives the average weighted duration between two neighbouring papers, where at least one of the papers has been published in the given year.

Figure 17: Topic Incidences (All Papers, Paths of Length 3)

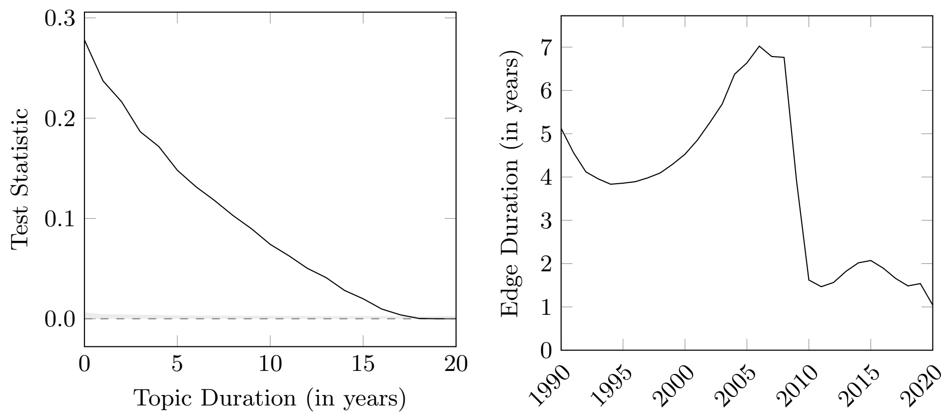


(a) Original Data (b) Bootstrapped Data

Notes: Each panel shows adjusted frequencies of edges between papers, tabulated by the years of those papers. Panel (a) shows the actual frequencies, whereas Panel (b) shows simulated frequencies when the years of papers are randomly perturbed.

Figure 18: Topic Duration (All Papers, Paths of Length 3)

(a) Significance Tests (b) Average Edge Duration



Notes: the left panel shows a test statistic for an $\mathcal{H}_0$ that topics are independent of publication year against an alternative $\mathcal{H}_a^t$ that some topics stay popular for exactly t years, where t is shown on the horizontal axis. The grey area shows the confidence interval where $\mathcal{H}_0$ cannot be rejected at 5%. The right panel gives the average weighted duration between two neighbouring papers, where at least one of the papers has been published in the given year.