

Is Dynamicity All You Need?

Richard Delwin Myloth, Kian Ahrabian, Arun Baalaji Sankar Ananthan, Xinwei Du, Jay Pujara

University of Southern California, Information Sciences Institute
{myloth, ahrabian, arunbaal, xinweidu, jpujara}@usc.edu

Abstract

Scientific domains are fluid entities that change and turn as time passes. Take machine learning as an example. Up until the '90s, most of the methods were expert-knowledge-driven. However, as time passed, more data-driven approaches appeared, finally leading to the advent of deep learning methods. As a result, in a span of 30 years, the field has gone through many changes and breakthroughs and is at a point where many novelties have a life span of shorter than five years. In parallel, a regular researcher's career span is around the same length. Consequently, being a researcher requires shifts in the field of study throughout one's career. Besides, researchers' scientific interests are inherently dynamic and change over time. Hence, there exists a dynamicity to authors' interests and fields of work over time. In this work, we study this phenomenon through systematic approaches for representing and tracking dynamicity in different epochs. Our representation approaches are based on the idea that each author could be represented as a distribution of other authors. Concurrently, our tracking approaches rely on established mathematical concepts for measuring the change between two distributions. We focus on the publications in the 2001-2020 range and present a set of analyses built on top of the introduced approaches to understanding the potential connection between dynamicity and success.

1 Introduction

The past few decades have been an unprecedented era of scientific discoveries, with the sheer number of publications rising steadily (Bornmann and Mutz 2015). This constant growth of research collaborations has led to the emergence of new interdisciplinary domains, prompting researchers to expand their research horizons. This expansion, combined with the continuous development of scientific domains and the inherent nature of research to explore new areas, results in a potentially volatile set of research directions. This work introduces approaches for systematically studying this fluidity and uncovering interesting behaviors among authors.

Scientific publications are the information vessels scientists use to communicate their findings, methodologies, and critiques. At the same time, publications are reflections of their authors' interests and fields of study. These publications are bound together through citations that specify the

foundations of each work. As a result, citations create tightly connected groups of publications with similar research directions. Consequently, authors with a high number of interactions in these groups, either through collaborations or citations, are more likely to have similar interests.

Community detection algorithms are graph partitioning approaches that identify sets of tightly connected nodes that are loosely connected to nodes outside their respective sets (Blondel et al. 2008; Traag, Waltman, and Van Eck 2019). When employed on citation networks, these algorithms yield a set of communities where each community contains highly related publications. These extracted communities could then be exploited for indirectly analyzing authors' interests through publications and citations as proxies.

In this work, we study the authors' dynamicity phenomenon from a relational standpoint. More specifically, we focus on the following research questions:

1. **How can we characterize and quantify the interests and dynamicity of an author?**
2. **Is there any connection between dynamicity and success due to reasons such as adaptability or diversity?**

To this end, we first create two knowledge graphs (KG) from publications in the 2001-2020 period, each encompassing ten years' worth of scholarly information, i.e., publications and authors. Then, we introduce three vectorizing approaches focused on presenting authors' interest in one epoch, and two tracking approaches focused on quantifying the change in interests in two distinct epochs. Our vectorizing approaches are built on top of relational information in the KGs and represent authors as a distribution of other authors. Meanwhile, our tracking approaches are based on the two well-known cosine similarity and relative entropy (Kullback–Leibler divergence) measures. By mix-and-matching, these approaches yield six different dynamicity scores for each author. We then use these scores to investigate the connection between authors' dynamicity and success. Our analyses showcase the connection between success, diversity, and adaptability in research.

2 Related Work

Bird et al. (2009) analyzed community structures in the DBLP bibliographic database to investigate collaborative

Dataset	CG-2010	CG-2020
# Publications	19,707,369	33,743,276
# Authors	20,333,216	36,077,559
# Citation Links	167,133,583	323,927,950
# Authorship Links	67,531,472	137,160,724

Table 1: Statistics of the extracted KGs.

connections in computer science and interdisciplinary research at the individual, within-area, and network-wide levels. They developed quantifiable metrics such as longitudinal assortativity over the number of publications, collaborators, and career length to study author overlap and migration patterns. Prior to Bird et al. (2009), Newman (2001) used data from publications in physics, biomedical research, and computer science to build co-authorship collaboration networks. They looked at the number of publications produced by authors, the number of authors per article, the number of collaborators that scientists have, the existence and size of a significant component of connected scientists, and the degree of clustering in the networks. They examined collaboration patterns among participants and discovered that these variables follow a power law distribution and that collaboration relationships are transitive. Paul et al. (2015) also used the DBLP database in their study to develop a citation-collaboration network to rank authors based on their contributions in terms of co-authorship and citations while verifying them against the h-index. They also carried out a comparative examination of the change in author ranking for different parts of the author spectrum over time.

3 Dataset

OpenAlex (Priem, Piwowar, and Orr 2022) is a free and open catalog of scholarly entities that provides metadata for publications, authors, venues, institutions, and scientific concepts, along with the relationships among them. It gathers data from sources such as Crossref, Microsoft Academic Graph (MAG), ROR, ORCID, DOAJ, PubMed, PubMed Central, and Unpaywall. We use the OpenAlex dump obtained on 2022-12-07 to construct our dataset for this work. Given this dump, we first extract a KG containing all the publications and their connections, i.e., citation links. Then, we extract two induced KGs by filtering the publications with publication dates within two ranges of 2001-2010 and 2011-2020, naming them CG-2010 and CG-2020, respectively. Following this, we add the authorship information for each KG for all the publications. Finally, we drop all the nodes with a zero degree (in and out) in both KGs. After this procedure, we end up with two temporally-scoped KGs containing authorship and citation information for all the publications in the 2001-2010 and 2011-2020 periods. Table 1 illustrates the statistics of the extracted KGs. To handle the large size of the raw dump, we resorted to using the KGTK toolkit for all our KG processing procedures (Ilievski et al. 2020).

4 Methodology

We break down the problem of characterizing authors’ dynamicity into two sets of approaches: **Vectorizers** and **Trackers**. Vectorizers, as described in Section 4.1, focus on presenting authors’ interest in one epoch. As described in Section 4.2, trackers focus on quantifying the change in interests in two distinct epochs. When combined, these approaches provide a systematic way of characterizing authors’ dynamicity.

4.1 Vectorizers

We introduce three approaches for vectorizing authors’ interests in a given epoch. The main idea of all these approaches is that each author’s interests could be modeled through a distribution over the set of other authors. Our first two approaches rely only on the information that could be directly extracted from citation links. In contrast, the third approach uses external information by building upon the output of a community detection algorithm. As a result, the third approach is prone to erroneous information propagated from the underlying community detection algorithm; in return, it gains access to more complex information compared to the first two approaches.

Co-authors In this approach, we present an author’s interests through their co-authors. To this end, given two arbitrary authors p and q and epoch t , we define the co-author weight value $\psi_p^t(q)$ as

$$\psi_p^t(q) = |\mathcal{V}_p^t \cap \mathcal{V}_q^t| \quad (1)$$

where \mathcal{V}_x^t is the set of publications by author x in epoch t . Building on top of these co-author weight values, for any arbitrary author p , we form the representative vector z_p^t as

$$z_p^t = [\psi_p^t(a_0), \psi_p^t(a_1), \dots, \psi_p^t(a_{|\mathcal{A}|})] \quad (2)$$

where \mathcal{A} is the set of all authors in the KG. It is important to note that these representative vectors are extremely sparse due to the large cardinality of \mathcal{A} .

Citations In this approach, we present an author’s interests through its citing and cited authors. To this end, given two arbitrary authors p and q and epoch t , we define the citation weight value $\phi_p^t(q)$ as

$$\phi_p^t(q) = \sum_{v \in \mathcal{V}_p^t} |\mathcal{N}_v^t \cap \mathcal{V}_q^t| + \sum_{u \in \mathcal{V}_q^t} |\mathcal{V}_p^t \cap \mathcal{N}_u^t| \quad (3)$$

where \mathcal{V}_x^t is the set of publications by author x in epoch t and \mathcal{N}_y^t is the set of all publications cited by publication y in epoch t . Building on these citation weight values, for any arbitrary author p , we form the representative vector z_p^t following Equation 2, replacing ψ_p^t with ϕ_p^t .

Communities In this approach, we present an author’s interests through authors with whom they publish in the same research communities. To this end, given a KG encompassing epoch t , we first extract the citation graph by removing all non-publication nodes, i.e., authors. Then, we run the Leiden (Traag, Waltman, and Van Eck 2019) community detection algorithm to extract a set of communities \mathcal{C} . We rely

on the hypothesis that each community represents a somewhat unique field of study. We use a modified version of the Leiden algorithm that limits the maximum number of generated communities and the number of publications in a community. Doing so avoids the creation of large unfocused, or small insignificant communities. Given the set of extracted communities \mathcal{C} , for any two arbitrary authors p and q , we define the co-occurrence weight value $\eta_p^C(q)$ as

$$\eta_p^C(q) = \begin{cases} \sum_{c \in \mathcal{C}} \frac{|c_p|}{|\mathcal{V}_p^t|} \log_2(|c_q| + \alpha) & p \neq q \\ 0 & p = q \end{cases} \quad (4)$$

where c_x is the set of publications by author x in community c , \mathcal{V}_x^t is the set of publications by author x in epoch t , and $\alpha = 0.001$. In this formalization, the effect of each community is weighed on the number of publications an author has in that community, e.g., $\frac{|c_p|}{|\mathcal{V}_p^t|}$. Moreover, each author’s influence is smoothened by taking the log value of their number of publications, e.g., $\log_2(c_q + \alpha)$. The resulting equation highlights the connection between any two authors that have many papers in the same communities and simultaneously waives the need for tracking the communities themselves. Building on top of these co-occurrence weight values, for any arbitrary author p , we can form a representative vector z_p^t following Equation 2, replacing ψ_p^t with η_p^C .

4.2 Trackers

We introduce two tracking approaches for quantifying the dynamicity between two distinct epochs. These two approaches are built on well-known mathematical concepts of cosine similarity and relative entropy.

Cosine Similarity (\mathcal{S} -score) Given the representative vectors of an arbitrary author p from two time periods, z_p^t and $z_p^{t'}$, we calculate the cosine similarity score $\mathcal{S}_p^{t,t'}$ defined as

$$\mathcal{S}_p^{t,t'} = \frac{z_p^t \cdot z_p^{t'}}{\|z_p^t\| \|z_p^{t'}\|}. \quad (5)$$

The calculated cosine similarity scores represent the stability of authors’ interests in two epochs, i.e., the higher the value, the more consistent the authors’ interests.

Relative Entropy (\mathcal{E} -score) Building on top of the representative vectors, for each arbitrary author p in period t , we define a probability distribution as

$$\mathcal{F}_p^t(q) = \frac{z_p^t[q] + \epsilon}{\sum_{q' \in \mathcal{A}} z_p^t[q'] + \epsilon |\mathcal{A}|} \quad \forall q \in \mathcal{A} \quad (6)$$

where $\epsilon = \frac{1}{|\mathcal{A}|}$ is the prior probability and \mathcal{A} is the set of all authors in the KG. Then, given the probability distributions of an arbitrary author p from two time periods, \mathcal{F}_p^t and $\mathcal{F}_p^{t'}$, we calculate the relative entropy $\mathcal{E}_q^{t,t'}$ as

$$\mathcal{E}_p^{t,t'} = D_{\text{KL}}(\mathcal{F}_p^{t'} \parallel \mathcal{F}_p^t) = \sum_{q \in \mathcal{A}} \mathcal{F}_p^{t'}(q) \log\left(\frac{\mathcal{F}_p^{t'}(q)}{\mathcal{F}_p^t(q)}\right). \quad (7)$$

In contrast to the cosine similarity score, the calculated relative entropy scores represent the volatility of authors’ interests in two epochs, i.e., the higher the value, the less consistent the authors’ interests are.

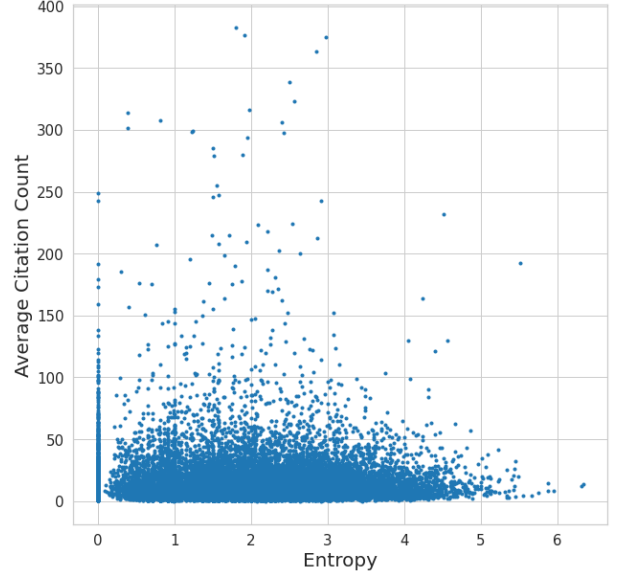


Figure 1: The effect of entropy on average citation count.

5 Analyses

Throughout this section, we run all our analyses on a set of randomly 10,000 sampled authors. More specifically, we do a weighted sampling without replacement using the citation counts. This procedure allows us to manage the computational costs of running these analyses.

5.1 Statistical Dependence Analysis

This analysis studies the connection between the introduced stability scores and success across two epochs. We use the relative change in average citation count as the proxy metric for success. The main intuitions behind this metric are 1) citation count is an accepted correlated metric for success in the community, 2) using average mitigates the effect of the high number of publications from an author, and 3) using relative change locally normalizes the metric values. Moreover, to reduce the potential noise in the data, we remove the outliers by filtering out samples outside two standard deviations of relative change in average citation count mean.

To quantify the strength of this connection, we use the established bivariate correlation and univariate linear regression measurements. We also include a random noise vectorizer as a sanity check to our methodology. Table 2 presents the results of our analysis with one of the introduced scores as the independent variable \mathcal{X} and the number of citations as the dependent variable \mathcal{Y} . As evident from Table 2, every introduced score has a significant connection with success, some in the same direction and some in the opposite direction. Moreover, the ‘‘Citations’’ vectorizer showcases the highest correlation with the measurement for success which signifies the effect of author interactions.

5.2 Entropy Analysis

In this analysis, we study the connection between diversity and success. We use the authors’ entropy across the ex-

Tracker	Vectorizer	PCC	Coef.	SE	t	$P > t $
\mathcal{S} -score	Random	-0.001	-967.70	5156.52	-0.188	0.851
	Co-authors	-0.121	-26.03	2.15	-12.11	0.000
	Citations	-0.138	-27.95	2.02	-13.81	0.000
	Communities	-0.082	-25.72	3.17	-8.12	0.000
\mathcal{E} -score	Random	0.015	47.03	31.15	1.51	0.131
	Co-authors	-0.057	-0.64	0.11	-5.65	0.000
	Citations	0.198	3.019	0.15	20.00	0.000
	Communities	0.048	0.66	0.14	4.73	0.000

Table 2: Univariate linear regression and bivariate correlation metrics between introduced scores and relative change in average citation count. Legend: **PCC**: Pearson correlation coefficient.

Metric	Est.	SE	z	$P > z $
ATE	-189.157	36.274	-5.215	0.000
ATT	-176.136	29.762	-5.918	0.000
ATU	-202.178	43.471	-4.651	0.000

Table 3: Treatment effect evaluations. Legend: **ATE**: Average treatment effect, **ATT**: Average treatment effect on the treated, **ATU**: Average treatment effect on the untreated.

tracted communities as a proxy for diversity. As for success, with similar intuitions to the previous section, we use the average citation count as the proxy metric. Formally, given the set of extracted communities \mathcal{C} , for any arbitrary author p , we calculate the entropy across communities $\mathcal{H}_p^{\mathcal{C}}$ as

$$w_p^c = \frac{|c_p|}{|\mathcal{V}_p^t|} \quad (8)$$

$$\mathcal{H}_p^{\mathcal{C}} = - \sum_{c \in \mathcal{C}} w_p^c \log_2(w_p^c) \quad (9)$$

where c_x is the set of publications by author x in community c and \mathcal{V}_x^t is the set of publications by author x in epoch t . Figure 1 illustrates the results of our analysis. We can observe in Figure 1 that in both epochs average citation count increases with the increase of entropy up until a point and then drops again. This observation indicates the benefit of having a diverse portfolio, but simultaneously too much diversity could negatively impact success.

5.3 Propensity Score Matching Analysis

This analysis focuses on the potential causal relationship between adaptability and success in two epochs by utilizing the propensity score matching (PSM) technique. We use the increase in entropy and citation count in the second epoch as proxy metrics for adaptability and success, respectively. Following this, we designate the increase in entropy as the treatment variable and the citation count in the second epoch as the outcome variable. As for the confounding variables, we use the publication counts from both epochs and the citation count in the first epoch. To check the matching quality, we plot one of the confounding variables, i.e., publication counts in the second epoch, against the outcome variable for

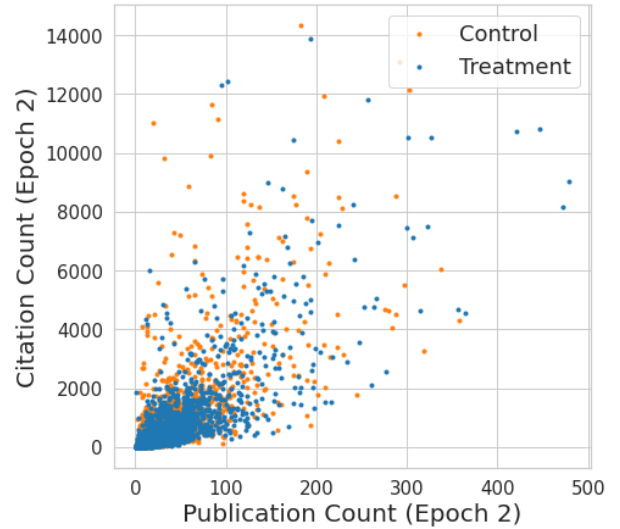


Figure 2: Matched groups for the confounding variable, i.e., publication count in the second epoch, for both control and treatment groups against the outcome variable.

both control and treatment groups in Figure 2. Moreover, Table 3 presents the treatment effect evaluation results. From Table 3, we can observe that the average treatment effect (ATE) has a larger value compared to the average treatment effect on treated (ATT) while both have a negative value. This observation indicates that while, in general, the authors have experienced a decline in the number of citations, the increase in entropy slows down this phenomenon. Hence, adaptability, i.e., an increase in entropy, could be seen as a remedy for a decline in success.

6 Conclusion and Future Works

Motivated by our observation of scientific domains' fluidity and empowered by the emergence of public repositories of scholarly data, we presented a thorough systematic study of the author dynamicity phenomenon in this work. With the idea of representing authors' interests and fields of work by a distribution of other authors, we introduced three different systematic approaches vectorizing each author in a single epoch. Then, to track an author's behavioral changes between two epochs, we introduced two approaches built on top of the extracted vectors and well-known mathematical approaches for quantifying change. Based on these approaches, we presented in-depth analyses to understand the connection between success better, as measured by citation counts, and specific dynamic behaviors, as measured through the introduced approaches.

Some of the straightforward extensions of our work for future studies are 1) including more authors, 2) using a more extended period, and 3) changing the temporal granularity for tracking changes. Moreover, we used a relatively simple metric as our success proxy; future works could work with other metrics, such as the h-index or i10-index.

Acknowledgements

This work was funded by the Defense Advanced Research Projects Agency with award W911NF-19-20271 and with support from a Keston Exploratory Research Award.

References

- Bird, C.; Barr, E. T.; Nash, A.; Devanbu, P. T.; Filkov, V.; and Su, Z. 2009. Structure and Dynamics of Research Collaboration in Computer Science. In *SDM*.
- Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10): P10008.
- Bornmann, L.; and Mutz, R. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11): 2215–2222.
- Ilievski, F.; Garijo, D.; Chalupsky, H.; Divvala, N. T.; Yao, Y.; Rogers, C.; Li, R.; Liu, J.; Singh, A.; Schwabe, D.; et al. 2020. KGTK: a toolkit for large knowledge graph manipulation and analysis. In *International Semantic Web Conference*, 278–293. Springer.
- Newman, M. E. 2001. Scientific collaboration networks. I. Network construction and fundamental results. *Phys Rev E Stat Nonlin Soft Matter Phys*, 64(1 Pt 2): 016131.
- Paul, P. S.; Kumar, V.; Choudhury, P.; and Nandi, S. 2015. Temporal analysis of author ranking using citation-collaboration network. In *2015 7th International Conference on Communication Systems and Networks (COMSNETS)*, 1–6.
- Priem, J.; Piwowar, H.; and Orr, R. 2022. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.
- Traag, V. A.; Waltman, L.; and Van Eck, N. J. 2019. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1): 1–12.