

Citations Driven by Social Connections? A Multi-Layer Representation of Coauthorship Networks

Vahan Nanumyan* Christian Zingg[†] Frank Schweitzer¹

Chair of Systems Design, ETH Zurich, Switzerland

**vahan@nanumyan.com*

[†]*czingg@ethz.ch*

¹*fschweitzer@ethz.ch*

Abstract

To what extent is the citation rate of new papers influenced by the past social relations of their authors? To answer this question, we present a data-driven analysis of nine different physics journals. Our analysis is based on a two-layer network representation constructed from two large-scale data sets, INSPIREHEP and APS. The social layer contains authors as nodes and coauthorship relations as links. This allows us to quantify the social relations of each author, prior to the publication of a new paper. The publication layer contains papers as nodes and citations between papers as links. This layer allows us to quantify scientific attention as measured by the change of the citation rate over time. We particularly study how this change depends on the social relations of their authors, prior to publication. We find that on average the maximum value of the citation rate is reached sooner for authors who either published more papers, or who had more coauthors in previous papers. We also find that for these authors the decay in the citation rate is faster, meaning that their papers are forgotten sooner.

Keywords: Collective Attention, Citation Rate, Two-Layer Network, Citation Network, Collaboration Network, Network Centrality

1 Introduction

The availability of large-scale data sets about journals and scientific publications therein, their authors, institutions, cited references and citations obtained in other papers has boosted scientometric research in the past years. They allow to address new research questions that go beyond the calculation of mere bibliographic indicators. These regard particularly the role of *social influences* on the success of papers,

for example coauthorship relations [19] or the relations between authors and handling editors [18]. Such investigations have contributed to a new scientific discipline, *the science of success* [10, 21].

But such data also allows to redo traditional scientometric analyses on a much larger scale. In [15], the dynamics of the citation *rate*, i.e. the change in the number of citations during a fixed time interval, is analysed. The authors find that the change of the average citation rate follows two characteristic phases, first a growth phase and then a decay phase. Interestingly, the duration of the first and the speed of the second phase have changed over the years. This allows conclusions of how the *collective attention* of scientists towards a given paper has evolved between early and recent times.

The recent progress on such questions very much relies on representing scientometric systems as networks. A first example are *citation networks* representing papers as nodes and citations as their (directed) links. Such networks can be seen as a knowledge map of science [11]. They can be also used to predict scientific success [13]. A second example are *coauthorship networks* representing scientists as nodes and their coauthorships as links. While sociological studies [3] just report that communication between coauthors can be very intricate, also formal models of how such collaborations form on the structural level have been developed [9, 23]. To study collaboration patterns in a university faculty [4], such coauthorship networks have been combined with a network encoding the physical distance between the faculty members. It was also analysed how communities detected on a coauthorship network overlap with different research topics [1].

These investigations have the drawback that they study citation networks and coauthorship networks separately from each other. As already emphasized [5, 20], this becomes a problem if one wants to study social influence on citation dynamics. For example, based on a data set of *Physical Review* it was shown that scientists cite former coauthors more often [12]. Therefore, a better approach is to combine both the citation and the coauthorship network in a *multilayer network*. Links between the citation and the coauthorship layer express the authorship of papers. Using such a representation, a method to detect citation cartels was proposed [7]. Further, the rate of citations dependent on the authors' total number of citations was studied [16]. However, it was not investigated yet how the *position* of authors in the coauthorship network influences *when* their papers are cited. In this paper we study exactly this question.

Our analysis extends recent studies that focus on the success of papers as measured by their total number of citations. In [19], this success was related to the position of the authors in a coauthorship network. It was shown that authors of successful papers are

considerably more central (as quantified by *various* centrality measures) in the coauthorship network. We extend this by an analysis of the dynamics of the citation rate over time, i.e. *when* their papers are cited. To parametrise the citation dynamics, we resort to the mentioned phases identified in [15]. We extend this work by relating these phases to the social relations of the authors.

Our paper is structured as follows. In Sect. 2.1 we explain how citation dynamics can be measured by means of *citation histories*, which represent the collective attention. In Sect. 2.2 we describe the data sets used for our analysis. In Sect. 3.1 we introduce the multilayer network to combine social information about authors with citation data. We then turn to our research question and study in Sects. 3.2, 3.3 how the social relations of authors in the coauthorship network influence the collective attention. Lastly, in Sect. 4 we conclude our findings.

2 Methods and Data

2.1 Dynamics of Citation Rates

Measuring attention. Citations are often used as a measure of *success* of a paper, accumulated over time. They have the advantage that they are objective in the sense that they are protocolled in the reference lists of citing papers. But the sheer number of citations does not utilize the *temporal information*, i.e. how many of these citations arrive at a given time. This is captured in the *citation rate*, which better estimates the *attention* a paper receives at a given time (interval). Individual attention, i.e. *who* cites a given paper at a given time, is not of interest for our study. We focus on *collective* attention, i.e. aggregate over all authors who cite this paper during a given time interval. Obviously, the citation rate is only a proxy of this collective attention. One could additionally consider other attention measures like the `altmetric` score. But such information is only available for very recent publications and further strongly biased against the use of social media. Therefore, we decide to restrict our study to only using the citation rate as a proxy for collective attention. At least, most papers are still cited because they have caught in some way the attention of the authors of the citing papers.

Citation histories. We measure the collective attention of a paper by the number of citations it receives over a particular time interval, i.e. its citation rate. More precisely, for paper i published at time δ_i , the citation rate at $t = \delta - \delta_i$ time units after publication

is

$$c_i(t) = \frac{k_i^{\text{in}}(\delta + \Delta t) - k_i^{\text{in}}(\delta)}{\Delta t} \quad (1)$$

where $k_i^{\text{in}}(\delta)$ denotes the total number of “incoming” citations the paper has received at time δ . The dynamics of the citation rate $c_i(t)$ is also called *citation history* of paper i [15]. To compare citation histories across papers we further normalise them by their respective maximum value $c_i^{\text{max}} = \max_t\{c_i(t)\}$:

$$\tilde{c}_i(t) = c_i(t)/c_i^{\text{max}}. \quad (2)$$

Two phases in citation histories. Parolo *et al.* [15] find two characteristic phases in the dynamics of normalised citation histories $\tilde{c}_i(t)$ of a paper i . In the first phase, which lasts for 2–7 years, it grows and eventually reaches a peak at a time t_i^{peak} . After the peak there is the second phase, in which the citation rate decays over time. For the majority of papers this decay was found to be well described by an exponential function:

$$\tilde{c}_i(t) \propto \exp(-t/\tau_i), \quad (3)$$

Figure 1 illustrates these two phases.

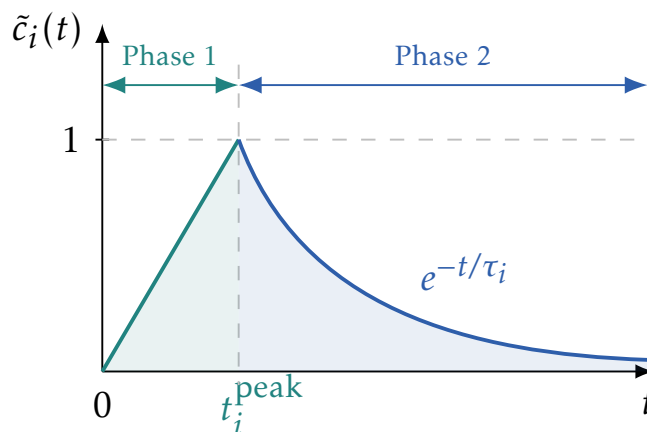


Figure 1: Illustration of the two characteristic phases in normalised citation histories $\tilde{c}_i(t)$ of most papers.

2.2 Bibliographic databases

To obtain the data for our study we resort to large bibliographic databases which index papers across journals. They collect information such as a paper’s title, the list of

authors, the date of publication, and also the list of references that a paper cites. We extracted this set of information from two such databases as explained in Appendix A. The first database indexes papers in journals by the American Physical Society (APS) and can be downloaded for research purposes from their website. From this database we extracted data for the journals Physical Review (PR), Physical Review A (PRA), Physical Review C (PRC), Physical Review E (PRE), and Reviews of Modern Physics RMP to cover a wide range of physics sub-fields. The second database, called INSPIRE-HEP, indexes papers relevant for the field of high-energy physics. From this database we extracted data for those journals with more than 10000 citations between papers in the respective journal. These journals are Journal of High Energy Physics (JHEP), Physics Letters (Phys. Lett.), Nuclear Physics (Nuc. Phys.), and high energy physics literature in Physical Review journals (PR-HEP).

Table 1: Overview of the extracted journals from the APS database and the INSPIRE-HEP database (IH). $|V^p|$ is the number of papers, $|V^a|$ is the number of authors, $|E^{pc}|$ is the number of citations between papers, and $|E^a|$ is the number of authorships.

Database	Journal	$ V^p $	$ V^a $	$ E^{pc} $	$ E^a $
APS	PR	46728	24307	253312	87386
	PRA	69147	41428	416639	144806
	PRC	36039	22672	253948	108844
	PRE	49118	36382	182701	95796
	RMP	3006	3788	5282	5044
IH	JHEP	15739	7994	191990	39056
	PR-HEP	44829	33908	213625	115237
	Phys. Lett.	22786	18078	56332	53089
	Nuc. Phys.	24014	18733	125252	60018

Table 1 provides summary statistics of the mentioned 9 journals. It further also shows how large these journals actually are. For example, there is only 1 journal, RMP, which contains less than 10000 authors, or there are more than 400000 citations between papers in PRA.

3 Social Influence on Citation Rate

3.1 Multilayer Network Representation

Combining information about papers and authors. Our aim is to combine the information about collective attention, as proxied by the citation rate, with information about the social relations between authors. For the latter, we specifically focus on *coauthorship*, because this is the most objective and best documented relation. Again, this is a proxy because it neglects other forms of social relationships, such as friendship, personal encounters, e.g. during conferences, electronic communication, or relations in social media. But we do not have this type of information available for all authors over long times. Therefore we restrict our analysis to the coauthorship network that can be constructed from available data, as described below.

To relate information about authorship and about papers in a tractable manner, multilayer networks come into play, because they allow us to represent such separate information in different layers. The nodes on the first layer correspond to papers and the (directed) links to their citations. Different from this, the nodes in the second layer correspond to the authors and the links to their coauthorships, i.e., there is a link between two authors if they wrote at least one paper together. Then, there are links which connect nodes on the first layer with nodes on the second layer. These links correspond to the authorship relations, i.e., for every author there is exactly one such link to each of her papers. We construct such a two-layer network for each of the 9 journals in our data set to represent the information about citations between papers as well as about the authorships.

To summarize the above, Figure 2 illustrates the two layers of citation and coauthorship networks and their coupling. It further displays the temporal dimension: The multilayer network evolves over time because new papers are published, and hence new coauthors appear. As the timeline indicates, paper i is published at time δ_i and then accumulates citations in the future, at times $\delta > \delta_i$. The publication layer allows us to define the degree of a *paper* i as the number $k_i^{\text{in}}(\delta)$ of papers that cite i until time δ , see Eq. (1) and Figure 2. Specifically, it is the *in-degree* because the publication network is directed. The question is now how the citation rate of this paper evolves over time, conditional on the social information about its authors at time δ_i , which is the publication time of paper i . In other words, we analyse the impact of information from *before* this publication.

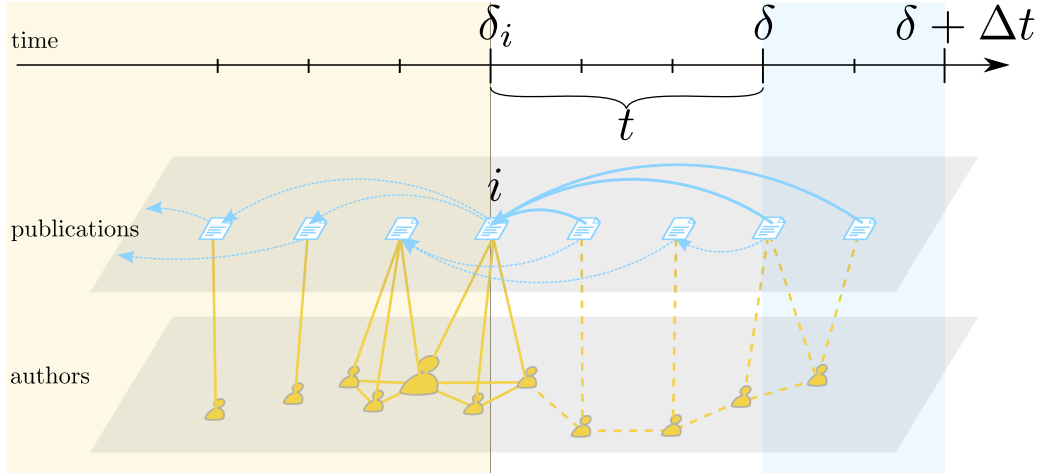


Figure 2: Multilayer network illustrating the coupling between the coauthorship network and the citation network. Links between the two layers represent the relation between authors and papers. The timeline on top indicates that links within the citation layer are directed and point to papers already existing at the time when a paper is published.

Quantifying authors' social relations. The coauthorship layer allows us to define the *degree* of an *author* n as the total number of distinct coauthors $k_n(\delta_i)$ that the author had *before* time δ_i . Degree is the simplest centrality measure for networks and reflects the *local* information about the embedding of an author in the social network. We use it here because it was shown recently [14] that this measure is a particularly good predictor for the future citation rate.

To characterize a paper i at time δ_i by using this information of its authors r *before* time δ_i , we sum up over their individual degrees k_r

$$s_i^{NC}(\delta_i) = \sum_r k_r(\delta_i) - C(\delta_i) \quad (4)$$

The index NC refers to number of coauthors. $C(\delta_i)$ is a correction term needed to only count unique coauthors (if some authors already share the same coauthor). This means, the paper i published at time δ_i is *not* counted in $s_i^{NC}(\delta_i)$.

We also make use of the coupling between the two layers for defining a second measure, which we can later compare with $s_i^{NC}(\delta_i)$. First we define the *interlayer-degree* $\tilde{k}_n(\delta_i)$ of an author n as the total number of distinct papers written by n *before* time δ_i . This measure allows us to quantify the experience of author n which she gained before a given point in time. To characterize a paper i at time δ_i by using this information of

its authors r before time δ_i , we compute

$$s_i^{NP}(\delta_i) = \sum_r \tilde{k}_r(\delta_i) - \tilde{C}(\delta_i) \quad (5)$$

in analogy to Eq. 4. Here $\tilde{C}(\delta_i)$ is again a correction term used to only count unique papers (if some authors had written a paper together in the past already).

Parametrising citation rates. The quantities $s_i^{NC}(\delta_i)$ and $s_i^{NP}(\delta_i)$ are based on the information of the *authors* of paper i . Our goal is to determine how they influence the *citation dynamics* of paper i . i.e. we need an analytically tractable parametrisation of the citation rates. To parametrise the citation dynamics we resort to the two characteristic phases of citation histories mentioned in Sect. 2.1. The first phase corresponds to increasing citation rates, and we parametrise by its duration t_i^{peak} , because we have no more precise knowledge about a general functional form of this phase. The second phase corresponds to an exponential decay, and we parametrise it as the parameter τ_i in Eq. 3, i.e. the so-called lifetime. Both parameters, t_i^{peak} and τ_i , are illustrated in Figure 1.

We now have four parameters to summarize the information about paper i . The first two parameters are $s_i^{NC}(\delta_i)$ and $s_i^{NP}(\delta_i)$, which characterise the *authors* of paper i . The other two parameters are t_i^{peak} and τ_i which characterise the *citation history*.

3.2 Time to the Peak Citation Rate

Regressions. Let us first analyse the relationship between the time t_i^{peak} until paper i reaches its highest citation rate, and the social relations of its authors. In order to find whether there is a significant relationship, we perform a linear analysis for log-transformed variables:

$$\log_{10} t_i^{\text{peak}} = \beta_0^{\text{peak}} + \beta_1^{\text{peak}} \cdot \log_{10} s_i \quad (6)$$

where s_i is the number of previous coauthors s_i^{NC} or the number of previous publications s_i^{NP} .

Results. They are presented in Table 2. All fitted parameters β_1^{peak} are negative and significantly different from 0 (on a significance level of 0.05). But how large is their effect for the untransformed variables? By exponentiating Eq. 6 we obtain the relation

$$t_i^{\text{peak}} \sim [s_i^{NC}]^{\beta_1^{\text{peak}}} \quad (7)$$

In particular, we see that β_1^{peak} becomes an exponent for the untransformed variables. Because it is negative, it follows that the time to reach the peak citation rate t_i^{peak} is *shorter* for papers with *more previous coauthors* (i.e. larger s_i^{NC}). For example, in the case of PRA the fitted $\beta_1^{\text{peak}} = -0.061$ predicts that the time it takes to reach the peak citation rate for a paper whose authors have 100 coauthors, will be on average 34% faster than for a paper whose authors only have 1 coauthor. We find the same statistically significant negative effect also for the number of previous publications s_i^{NP} . This means that the time it takes to reach the peak citation rate also tends to be *shorter* if the authors of the given paper have written *more publications* prior to it.

Table 2: Fitted citation rate parameters β_1^{peak} and β_1^τ for the examined journals in our INSPIREHEP data set (left) and in our APS data set (right). The parameters are the respective coefficients in the linear regressions in Equations (6) and (8). The significance levels of the p-values for the estimated parameters are encoded as *** (< 0.001), ** (< 0.01), * (< 0.05).

s_i	Time	β_1^{peak}	β_1^τ	s_i	Time	β_1^{peak}	β_1^τ
JHEP				PR			
NC	days	-0.048 ± 0.012 ***	-0.061 ± 0.009 ***	NC	days	-0.069 ± 0.008 ***	-0.082 ± 0.008 ***
	pubs	-0.048 ± 0.011 ***	-0.002 ± 0.008		pubs	-0.040 ± 0.007 ***	-0.013 ± 0.008
NP	days	-0.027 ± 0.010 **	-0.026 ± 0.008 ***	NP	days	-0.017 ± 0.008 *	-0.032 ± 0.008 ***
	pubs	-0.034 ± 0.010 ***	0.011 ± 0.007		pubs	-0.023 ± 0.007 ***	-0.020 ± 0.008 *
PR-HEP				PRA			
NC	days	-0.112 ± 0.007 ***	-0.159 ± 0.007 ***	NC	days	-0.061 ± 0.005 ***	-0.138 ± 0.005 ***
	pubs	-0.007 ± 0.006	-0.054 ± 0.008 ***		pubs	-0.018 ± 0.005 ***	-0.056 ± 0.005 ***
NP	days	-0.103 ± 0.006 ***	-0.135 ± 0.007 ***	NP	days	-0.063 ± 0.005 ***	-0.142 ± 0.006 ***
	pubs	-0.013 ± 0.006 *	-0.039 ± 0.007 ***		pubs	-0.022 ± 0.005 ***	-0.058 ± 0.005 ***
Phys. Lett.				PRC			
NC	days	-0.065 ± 0.012 ***	-0.080 ± 0.016 ***	NC	days	-0.019 ± 0.006 **	-0.071 ± 0.007 ***
	pubs	-0.084 ± 0.012 ***	-0.123 ± 0.016 ***		pubs	-0.007 ± 0.006	-0.052 ± 0.007 ***
NP	days	-0.069 ± 0.011 ***	-0.070 ± 0.014 ***	NP	days	-0.030 ± 0.007 ***	-0.083 ± 0.009 ***
	pubs	-0.083 ± 0.012 ***	-0.102 ± 0.014 ***		pubs	-0.021 ± 0.007 **	-0.063 ± 0.009 ***
Nuc. Phys.				PRE			
NC	days	-0.062 ± 0.009 ***	-0.117 ± 0.010 ***	NC	days	-0.023 ± 0.006 ***	-0.041 ± 0.007 ***
	pubs	-0.008 ± 0.009	-0.116 ± 0.010 ***		pubs	-0.014 ± 0.006 *	-0.049 ± 0.008 ***
NP	days	-0.069 ± 0.008 ***	-0.119 ± 0.008 ***	NP	days	-0.028 ± 0.006 ***	-0.036 ± 0.008 ***
	pubs	-0.000 ± 0.008	-0.111 ± 0.009 ***		pubs	-0.019 ± 0.006 **	-0.038 ± 0.009 ***
RMP				RMP			
NC	days	-0.106 ± 0.032 **	-0.272 ± 0.054 ***	NC	days	-0.106 ± 0.032 **	-0.272 ± 0.054 ***
	pubs	-0.126 ± 0.033 ***	-0.337 ± 0.051 ***		pubs	-0.126 ± 0.033 ***	-0.337 ± 0.051 ***
NP	days	-0.099 ± 0.036 **	-0.247 ± 0.070 ***	NP	days	-0.099 ± 0.036 **	-0.247 ± 0.070 ***
	pubs	-0.110 ± 0.036 **	-0.281 ± 0.062 ***		pubs	-0.110 ± 0.036 **	-0.281 ± 0.062 ***

3.3 Characteristic Decay Time

Regressions. To investigate the characteristic decay time, we again perform a linear analysis for log-transformed variables. This corresponds to the following model:

$$\log_{10} \tau_i = \beta_0^\tau + \beta_1^\tau \cdot \log_{10} s_i \quad (8)$$

where again s_i is the number of previous coauthors s_i^{NC} or the number of previous publications s_i^{NP} .

Results. They are presented in Table 2. Also here all fitted parameters β_1^τ are negative and significantly different from 0 (on a significance level of 0.05). This means that the more previous coauthors the authors have, the smaller the value of τ_i becomes. According to Eq. 3 this means that the more coauthors an author had before the current publication, the steeper the decay in the citation rate becomes. This in turn means that such a paper faces a quicker and stronger shortage in new citations. Again, we also find significantly negative parameters β_1^τ when using the number of previous publications s_i^{NP} in Equation 8.

3.4 Rescaling Time by Counting Publications

Effect of the growing scientific output. It is known that the number of papers published every year grows exponentially over time [17]. This means that in recent years there are more papers published in a given time-interval than this was the case longer ago. And all of these new publications can potentially cite a given paper. This time-dependence likely affects our regression results by confounding the respective response (t_i^{peak} or τ_i) and predictor variable (s_i^{NC} or s_i^{NP}). In the past it was suggested that the dependence of the citation rate on the publication year of a paper can be weakened by counting time in terms of the number of published papers instead of days, or weeks, etc. [15]. Therefore we repeat our regressions from Sects. 3.2, and 3.3 while measuring time on this alternative timescale. Thereby we assess whether such a bias from the publication year of a paper is present in the relations which we found.

Results for the alternative timescale. These fitted parameters are also listed in Table 2. For all but one journal they remain smaller than 0, just like before. The exception is JHEP whose parameter β_1^τ becomes positive when measured for s_i^{NP} as independent variable. However, this β_1^τ is not significantly different from 0 anymore on a significance level of 0.05, and hence does not contradict our previous findings. Overall, the

parameter β_1^{peak} is significant in all journals except PR-HEP, Nuc. Phys. and PRC and the parameter β_1^τ is significant in all journals except JHEP and PR. Hence we conclude that time does not introduce a general bias to our findings. This means that, also according to the alternative timescale, the peak in the citation rate is reached faster for papers by authors with more previous coauthors or publications. Accordingly, the decay becomes steeper for papers by such authors.

4 Conclusions

In this paper, we address the question how the attention towards an academic publication is accumulated over time, depending on the social relations of its authors, as expressed in the coauthorship network. For example, does the attention mostly occur in an early phase right after publication? Or is it rather spread uniformly over time? Or might it even happen only after a long time has passed since publication? To obtain a tractable, objective characterisation of attention, we proxy attention by the citation rate of a paper, i.e. the number of new citations obtained in a particular time-interval. We argue that, in order for a citation to occur, the authors of the citing paper have to be aware of the cited paper.

To study the time when this attention occurs, we compute the change in the number of citations over a time-interval, i.e. the citation rate. It is known that citation rates of most papers have two characteristic phases over time, namely an increasing phase followed by a decay phase. We found that the first phase tends to get shorter and the decay in the second phase tends to get faster for papers written by authors who have many previous coauthors. In terms of attention, our findings mean that such papers attract attention faster, but are then also forgotten sooner. We also found this effect when measuring the number of previous papers of the authors instead of the number of previous coauthors. Furthermore, this effect also persisted when we controlled for the time when a paper was published. But most importantly, we found this effect in 9 journals, based on *hundreds of thousands of authors and papers* and *far more than a million of citations*. A study of such a large scale is a strong sign that we uncovered a general trend which is not limited to the analysed data sets.

A speculative explanation. Which mechanisms could be responsible for this? One way how authors learn about the papers which they cite is through communication with other scientists. Hence, authors can use their (few or many) social contacts, proxied by coauthors, to “advertise” a paper. Our findings indicate that authors with many

previous coauthors or papers tend to do so within a short period of time after publication. When a new publication is made, the authors “advertise” it to the scientific community by presenting it in conferences and seminars, by sharing it on social media, etc. This behaviour happens within a finite time period, after which the authors stop actively promoting the given publication. However, this explanation is merely speculative at this point.

Regressions not suitable for predictions. Our performed regressions have low predictive power, as indicated by extremely small coefficients of determination, R^2 . For instance, for some regressions the R^2 is as low as 0.001, meaning that only 0.1% of the variance in the dependent variable is explained. However, while our regression models are not useful for prediction, our inferred relations are significant. In particular our regressions show that the time to the peak citation rate and the subsequent decay are not independent of the authors.

Future work. Our analysis shows that the social relations of authors significantly influence the citation histories of their papers. But our analysis does not provide insights about the *mechanisms* behind this influence. In the future, we can use generative modelling to learn more about these underlying mechanisms. For instance, hypotheses can be formulated and tested using the framework of coupled growth models presented in [14].

We find that a paper receives attention from the scientific community faster, the more coauthors the authors had prior to its publication. But we find as well that such a paper is then also forgotten sooner again. Our findings indeed highlight that the citations of a paper can have substantially different dynamics depending on the social relations of the authors. Our approach also illustrates how such coupled dynamics can be studied by representing scientific collaborations as a multilayer network.

References

- [1] Battiston, F.; Iacovacci, J.; Nicosia, V.; Bianconi, G.; Latora, V. (2016). Emergence of Multiplex Communities in Collaboration Networks. *PLoS ONE* **11(1)**, e0147451.
- [2] Burrell, Q. L. (2005). Are “sleeping beauties” to be expected? *Scientometrics* **65(3)**, 381–389.

- [3] Cetina, K. (2009). *Epistemic Cultures: How the Sciences Make Knowledge*. Harvard University Press.
- [4] Claudel, M.; Massaro, E.; Santi, P.; Murray, F.; Ratti, C. (2017). An exploration of collaborative scientific production at MIT through spatial organization and institutional affiliation. *PLOS ONE* **12(6)**, e0179334.
- [5] Clauset, A.; Larremore, D. B.; Sinatra, R. (2017). Data-driven predictions in the science of science. *Science* **355(6324)**, 477–480.
- [6] Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science and Technology* **52(7)**, 558–569.
- [7] Fister, I.; Fister, I.; Perc, M. (2016). Toward the Discovery of Citation Cartels in Citation Networks. *Frontiers in Physics* **4**, 49.
- [8] Frost, J. (2014). How Important Are Normal Residuals in Regression Analysis? Accessed: 2018-03-10.
- [9] Guimera, R. (2005). Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance. *Science* **308(5722)**, 697–702.
- [10] Jadidi, M.; Karimi, F.; Lietz, H.; Wagner, C. (2018). Gender disparities in science? Dropout, productivity, collaborations and success of male and female computer scientists. *Advances in Complex Systems* **21(03n04)**, 1750011.
- [11] Leydesdorff, L.; Carley, S.; Rafols, I. (2013). Global maps of science based on the new Web-of-Science categories. *Scientometrics* **94(2)**, 589–593.
- [12] Martin, T.; Ball, B.; Karrer, B.; Newman, M. E. J. (2013). Coauthorship and citation patterns in the Physical Review. *Physical Review E* **88(1)**, 012814.
- [13] Mazloumian, A. (2012). Predicting Scholars' Scientific Impact. *PLoS ONE* **7(11)**, e49246.
- [14] Nanumyan, V.; Gote, C.; Schweitzer, F. (N.D.). A multi-layer network approach to modelling authorship influence on citation dynamics in physics journals.
- [15] Parolo, P. D. B.; Pan, R. K.; Ghosh, R.; Huberman, B. A.; Kaski, K.; Fortunato, S. (2015). Attention decay in science. *Journal of Informetrics* **9(4)**, 734 – 745.

- [16] Petersen, A. M.; Fortunato, S.; Pan, R. K.; Kaski, K.; Penner, O.; Rungi, A.; Riccaboni, M.; Stanley, H. E.; Pammolli, F. (2014). Reputation and impact in academic careers. *Proceedings of the National Academy of Sciences* **111**(43), 15316–15321.
- [17] Price, D. J. (1951). Quantitative Measures of the Development of Science. *Archives Internationales d’Histoire des Sciences* **4**(14), 85–93.
- [18] Sarigol, E.; Garcia, D.; Scholtes, I.; Schweitzer, F. (2017). Quantifying the effect of editor-author relations on manuscript handling times. *Scientometrics* **113**(1), 609–631.
- [19] Sarigol, E.; Pfitzner, R.; Scholtes, I.; Garas, A.; Schweitzer, F. (2014). Predicting Scientific Success Based on Coauthorship Networks. *EPJ Data Science* **3**, 9.
- [20] Schweitzer, F. (2014). Scientific networks and success in science. *EPJ Data Science* **3**(1), 35.
- [21] Sinatra, R.; Lambiotte, R. (2018). Editorial. *Advances in Complex Systems* **21**(03n04), 1802001.
- [22] Sinatra, R.; Wang, D.; Deville, P.; Song, C.; Barabási, A.-L. (2016). Quantifying the evolution of individual scientific impact. *Science* **354**(6312).
- [23] Tomasello, M. V.; Vaccario, G.; Schweitzer, F. (2017). Data-driven modeling of collaboration networks: a cross-domain analysis. *EPJ Data Science* **6**(1), 22.
- [24] Van Raan, A. F. (2004). Sleeping beauties in science. *Scientometrics* **59**(3), 467–472.

A Data Sets

The data sets we use are obtained in the same way as in Nanumyan *et al.* [14]. In particular, we have 9 data sets which correspond to physics journals. We extracted these data sets from two large bibliometric databases which index academic papers, their citations, authorships, etc.

APS database. It indexes papers published in journals by the American Physical Society (APS). A dump of the database can be downloaded for research purposes at <https://journals.aps.org/datasets>. This database indexes authors by their name,

and hence different authors who share the same name are indexed as one author. For our study this has the drawback that such an author then has the publications and coauthors that were actually accumulated by multiple authors. In other words, when an author with a common name writes a paper i , our measures s_i^{NC} and s_i^{NP} are too large, i.e. biased. Or it can be that the same author is indexed with different name-spellings, and hence counts as multiple authors. For papers written by such an author s_i^{NC} and s_i^{NP} are too small, because not all previous coauthors or publications are actually counted.

To remove this bias we applied the disambiguation algorithm by Sinatra *et al.* [22] which was developed for the APS data set. After this name resolution we extracted the journals Physical Review (PR), Physical Review A (PRA), Physical Review C (PRC), Physical Review E (PRE), and Reviews of Modern Physics RMP to cover a wide range of physics sub-fields.

INSPIREHEP database. It indexes papers in the field of high-energy physics. A dump of the database can be downloaded at <http://inspirehep.net/dumps/inspire-dump.html>. This database has the advantage that the contained authors are disambiguated, meaning that two different authors with the same name are listed as separate authors. A drawback for our study is that some indexed papers have exceptionally large lists of authors which sometimes even exceed 1000 authors. Such large-scale collaborations are also called *hyperauthorships*, and the concern was raised that it is unclear what authorship actually means for such collaborations and which authors made substantial contributions [6].

We exclude such papers from our analysis. Our approach is based on our observation that these hyperauthorships occur mostly in papers related to large experiments, which are identified in the INSPIREHEP database by meta-tags. In particular, we considered only papers that are tagged as theoretical work or as a work in general physics. From these we then excluded papers that are in addition tagged as experimental work. After this filtering step we chose the largest 4 journals in terms of the number of citations that originate and end at a paper in the same journal. These are the Journal of High Energy Physics (JHEP), Physics Letters (Phys. Lett.), Nuclear Physics (Nuc. Phys.), and a collection of high energy physics literature in Physical Review journals (PR-HEP).

Excluding incomplete citation histories. Obviously our data sets only contain papers published before a certain point in time, namely the release date of the respective

database version. Therefore this release date also limits the time-span on which we can compute a given paper’s citation history. Especially for recent papers this introduces a problem: the observable part of the citation history might be so short that the decay phase has not yet started at all. This is in general difficult to detect, i.e. whether a particular citation history is covered over a sufficiently long period in a data set. To still account for this bias, we only consider citation histories of papers that were published 5 years before the end of a data set or earlier. Excluding papers from an even longer time period could have reduced this bias even more, but thereby we would also lose a considerable number of papers whose citation histories are likely already suitably represented. For instance, the currently 5 discarded years reduce the number of papers in PRA from 69147 to 54782, which is a decrease by 20%.

Besides excluding papers published within the last 5 years before the release of the respective database, we also exclude papers whose citation rate is growing at the latest observed time step. That is, we only consider papers which are already in the decay phase. This approach mostly eliminates papers published closer to the end of the data set. However, it also eliminates older papers that are called “sleeping beauties”—papers that remain unnoticed for a prolonged period of time, only to become frequently cited afterwards [2, 24]. However, sleeping beauties are extremely rare, so discarding them will not affect our statistical outcomes. For instance, Van Raan [24] identified only 0.04% of the articles published in 1988 as sleeping beauties.

B Regression Model Validations

B.1 Peak Delay Models

Aspects to validate. We show a validation here for the example of PRA. For a linear regression to be valid, the residuals must be normally distributed, their variance must not depend on the explanatory variables, and their expectation must be zero. To confirm that our regressions are valid, let us look at established diagnostic tools for these assumptions.

QQ-plot. To test for normality, we look at the Quantile-Quantile (QQ) plot between the observed distribution of the residuals and a theoretical normal distribution. If the observed distribution is the same as the theoretical one, the points in the QQ-plot will all fall close to the identity line. The result for the regression in Equation 6, $s_i = s_i^{NC}$ and time measured on the alternative time-scale is shown in the left panel of Figure 3. We

see that in the lower tail, i.e. for the smallest negative residuals, the lowest observed quantile stretches over almost the whole negative range of the theoretical quantiles. This is due to the finite size of the time unit over which we have computed the citation rates. Yet still, the plot indicates a violation of the normality assumption in our case. However, non-normality of the residuals means that a regression is not reliable for predictions, but it does not threaten the significance of the slope [8]. And the latter is the use-case in our study.

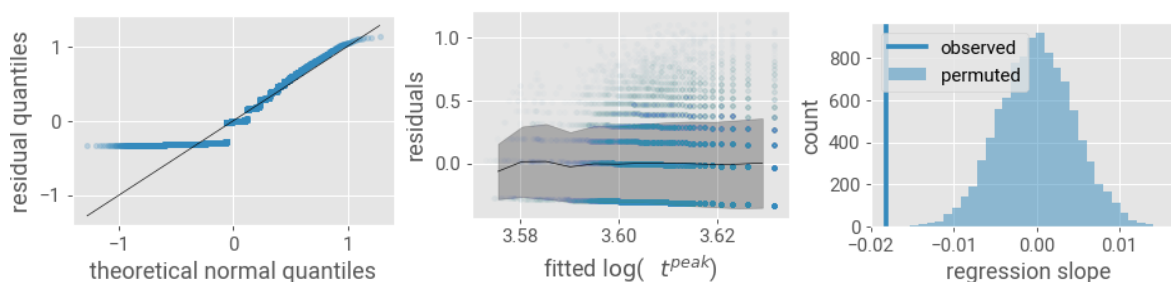


Figure 3: Validity of Equation 6 for PRA. (Left) quantile-quantile plot of residuals versus normal distribution, (middle) Tukey-Anscombe plot and (right) the result of permutation test based on resampling the data 10000 times.

Tukey-Anscombe plot. Next, we check whether the expectation of the residuals is zero and is independent from the explanatory variable. In the middle panel of Figure 3 we present the Tukey-Anscombe plot for the regression in Equation 6, $s_i = s_i^{NC}$ and time measured on the alternative time-scale. This plot shows the residuals plotted against the predicted values of the dependent variable. The black line is the mean of the residuals for different values of the dependent variable. We see that it is close to zero for all values of the predicted dependent variable.

Based on the Tukey-Anscombe plot, we can also check the condition of homoskedasticity, i.e., that the variances of the residuals are constant. To this end, we include the standard error of the residuals against the predicted value of the dependent variable. We see that it grows slightly with the dependent variable for small values. However, we also have to consider that the estimation of standard errors in that region is not very reliable due to the rather few observed data points. We conclude that overall the assumption of constant variance is well satisfied.

A cross-check. The mentioned plots allow only a qualitative evaluation of the validity of the regressions. As we are interested in the statistical significance of the dependence between the dependent and explanatory variables, we confirm the significance

of the slope by means of a permutation test. That is, we randomly reshuffle the values of the explanatory variable between different papers, while keeping the original values of the dependent variable. Then we perform the regression analysis on this shuffled data. By repeating this procedure multiple times, we obtain a distribution of values of the regression analysis based on the randomised data. The right panel of Figure 3 shows the outcome for the slope of the regression for 10000 randomised trials. We find that the distribution of the slope β_1^{peak} is centred around zero, meaning that there is no randomly expected effect merely from how the explanatory or dependent variables are distributed. We also find that the estimated slope from the observed data is far outside the distribution of the randomised data. This confirms the statistical significance of the identified dependence between the time to the peak citation rate and the number of previous coauthors of the authors. Hence, we can conclude that the social relations of authors have a significant influence on the time t^{extpeak}_i it takes a paper i to reach the largest citation rate.

B.2 Decay Models

Let us now check the validity of the linear regression analysis for one example, namely the regression in Equation 8, $s_i = s_i^{NC}$ and time measured on the alternative time-scale.

QQ-plot. In the left panel of Figure 4 we present the QQ-plot for the distribution of the residuals against the theoretical normal distribution. We see that the observed positive residuals tend to be slightly larger than demanded by the normal distribution. In contrast, the observed negative residuals, below one standard deviation, tend to be smaller than predicted. But overall, within one standard deviation, the observed distribution is reasonably similar to the normal distribution.

Tukey-Anscombe plot. Next, we inspect the Tukey-Anscombe plot shown in the middle panel of Figure 4. We find that the means of the residuals are very close to zero for all values of the predicted dependent variable, with the standard deviation being almost constant. We can conclude that the conditions of the validity of the linear regression are reasonably met.

A cross-check. To confirm the statistical significance that there is a dependence between the dependent and explanatory variables, we again also perform a permutation test by randomly reshuffling the values of the explanatory variable s_i^{NC} between

papers. Performing the regression on 10000 such randomisations, we obtain the distribution of the slope β_1^τ shown in the right panel of Figure 4. We see that the slope inferred for the empirical data is far from the distribution from the randomised data. Hence, we can conclude that the number of previous coauthors have a significant influence on the characteristic decay time of a paper. I.e. the more coauthors given authors have, the shorter is the characteristic decay time.

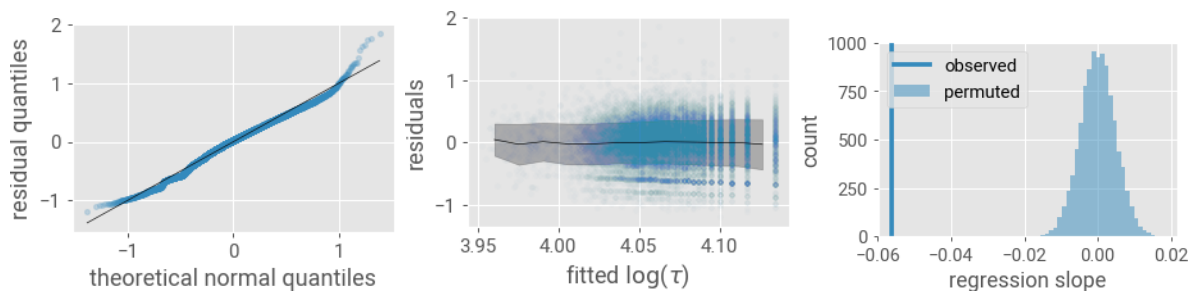


Figure 4: Validity of Equation 8 for PRA. (Left) quantile-quantile plot of residuals versus normal distribution, (middle) Tukey-Anscombe plot and (right) the result of permutation test based on resampling the data 10000 times.