Supplementary Material: Citations driven by social connections? A multi-layer representation of coauthorship networks

Christian Zingg[1], Vahan Nanumyan[1], and Frank Schweitzer[1]

[1] *Chair of Systems Design, ETH Zurich, Zurich, Switzerland*

S1 Correction Terms

$s_i^{NC}(\delta_i)$ is defined in Eq. (4) as the number of persons on the coauthorship layer who had co-authored at least one publication with any of the authors of paper i before time δ_i. The aim of the correction-term $C(\delta_i)$ in Eq. (4) is to correct the sum of the authors' degrees by those co-authors who collaborated with more than just one author. I.e., so that $s_i^{NC}(\delta_i)$ does not count some co-authors multiple times. Precisely, $C(\delta_i)$ can be computed as

$$C(\delta_i) = \max\left\{0, \sum_{s\in V^a}\left(\sum_r 1_{s,r}(\delta_i) - 1\right)\right\} \tag{S1}$$

where s is any person on the coauthorship layer, r is an author of paper i, and $1_{s,r}(\delta_i)$ is equal to 1 exactly if s and r have coauthored at least one paper before time δ_i, and is 0 otherwise. The intuition of Eq. (S1) is as follows. The sum over r counts for a specific person s in the coauthorship layer with how many of the authors of paper i she or he has at least one previous coauthorship before time δ_i. This sum is exactly equal to the number of times the sum of the degrees in Eq. (4) counts person r in total. But now we have to subtract 1 from this number, because otherwise we are removing person s completely from the sum of the degrees, whereas she or he should be contained with a weight of exactly 1. The "$\max\{0,\dots\}$" is a numerical trick to handle those persons s correctly who did not collaborate with any of the authors of paper i before time δ_i. If we would omit this part, they would add a negative weight to the correction term, instead of 0. Thereby, Eq. (S1) yields the correction necessary to adjust Eq. (4) into the number of persons on the co-authorship layer who have collaborated with at least one author before time δ_i.

In analogy to Eq. (S1) the correction term for Eq. (5) can be computed as

$$\tilde{C}(\delta_i) = \max\left\{0, \sum_{p\in V^p}\left(\sum_r 1_{p,r}(\tau_i) - 1\right)\right\} \tag{S2}$$

where p is any paper on the citation layer published before time δ_i, r is an author of paper i, and $1_{p,r}(\delta_i)$ is equal to 1 exactly if r is an author of paper p.

S2 Linear Regression Model Validation

Aspects to validate. 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 according to Eq. (7) are valid, we look at established diagnostic tools for these assumptions. We show a validation here for the example of the journal PRA, s_i as the number of previous coauthors s_i^{NC}, and time measured in years.

QQ-plot. It is presented in the left panel of Figure S1. We see that overall the points coincide well with the diagonal. Only in the tails, above 1 standard deviation and below -1 standard deviation, the residual quantiles tend to be slightly larger than those by the theoretical normal distribution. However, in these regions there are also only few observed points. Hence, we conclude that overall the vast majority of residuals is adequately described by a normal distribution.

Tukey-Anscombe plot. We also inspect the Tukey-Anscombe plot, which is shown in the right panel of Figure S1. We find that the means of the residuals are very close to zero for all values of the predicted decay exponents. Furthermore, also the standard deviation is almost constant for different fitted decay exponents. We conclude that all assumptions of linear regression are reasonably met for the regression in Eq. (7).

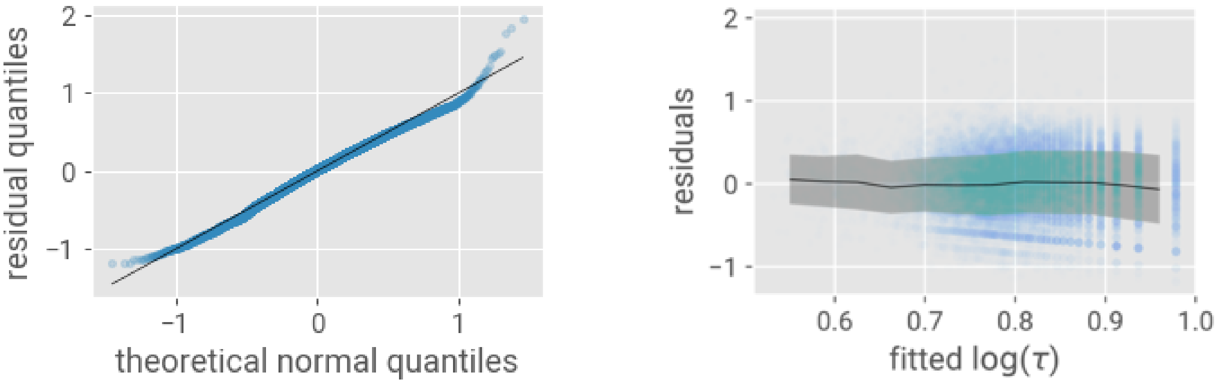

Figure S1: Validity of Eq. (7) for PRA. (Left) quantile-quantile plot of residuals versus normal distribution. (Right) Tukey-Anscombe plot.

S3 Testing for Over-Dispersion

In this Section, we validate that the variance of our peak-delays, t_i^{peak}, is larger than their mean. For count data, this condition implies that negative binomial regression shall be preferred over classical Poisson regression. To test this condition, we test the Null Hypothesis

$$H_0 : Var[t_i^{\mathrm{peak}} \,|\, s_i] = E[t_i^{\mathrm{peak}} \,|\, s_i] \tag{S3}$$

against the Alternative Hypothesis

$$H_A : Var[t_i^{\mathrm{peak}} \,|\, s_i] > E[t_i^{\mathrm{peak}} \,|\, s_i] \tag{S4}$$

To test these Hypotheses, we apply the over-dispersion test introduced in (Cameron & Trivedi, 1990), see their Section 2. To compute this test, we apply the function *dispersiontest* from the R-package *AER*. The obtained p-values are displayed in Table S1. We find that they are all highly significant, implying that H_0 should be rejected in favour of H_A in all of our journals. Hence, we choose negative binomial regression as our model for peak-delays.

Table S1: The p-values obtained for the over-dispersion test from (Cameron & Trivedi, 1990), computed for each journal and s_i separately. *NC* means that s_i^{NC} is used as s_i in Eqs. (S3), (S4). Accordingly, *NP* means that s_i^{NP} is used as s_i in these Eqs.

	s_i	p-value	
PR	NC	2.85×10^{-104}	***
	NP	5.72×10^{-94}	***
PRA	NC	1.80×10^{-191}	***
	NP	5.34×10^{-193}	***
PRC	NC	1.07×10^{-130}	***
	NP	4.51×10^{-129}	***
PRE	NC	1.51×10^{-164}	***
	NP	9.21×10^{-168}	***
RMP	NC	6.04×10^{-8}	***
	NP	1.37×10^{-7}	***

	s_i	p-value	
JHEP	NC	1.11×10^{-83}	***
	NP	3.18×10^{-83}	***
PR-HEP	NC	1.13×10^{-113}	***
	NP	6.68×10^{-117}	***
Phys. Lett.	NC	2.29×10^{-30}	***
	NP	1.14×10^{-30}	***
Nuc. Phys.	NC	2.90×10^{-94}	***
	NP	8.45×10^{-94}	***

References

Cameron, A., & Trivedi, P. K. (1990). Regression-based tests for overdispersion in the Poisson model. *Journal of Econometrics*, *46*(3), 347–364. doi: 10.1016/0304-4076(90)90014-K