Supplementary Material: Inferring the causal effect of journals on citations
V.A. Traag1, a)
Centre for Science and Technology Studies (CWTS), Leiden University, the Netherlands

I. DATA

We combined data from arXiv, Crossref and Scopus to establish our dataset. All data necessary to reproduce the results in this analysis is available from Traag (2020a) and all source code is available from Traag (2020b).

A. arXiv


For all arXiv XML elements in the data we extracted the arXiv identifier, and if present the DOI. We also extracted the date the preprint was first posted on arXiv. In total, this dataset covered 1,341,016 preprints, and a DOI is provided for 727,186 preprints (54%).

We extracted the subject for each arXiv preprint. The subjects were quite noisy, and did not contain only the subject division of arXiv, but also other subject classifications, most notably, the Mathematical Subject Classification (MSC). The arXiv subject classifications were provided as “Major - Minor” subjects, although sometimes only a major subject was provided. We extracted the major part and assigned an arXiv preprint to a major subject if that subject is at least used by 1,000 preprints (and is not an MSC). We thus retain 18 major subjects.

Preprints can be assigned to multiple major subjects. The large majority of arXiv preprints is assigned to a single major subject (80%). A single preprint has been assigned to as many as 8 different major subjects (1108.2700). There are only 261 preprints that have not been assigned to any of the major subjects. These are papers that are published in economics (33) and electrical engineering (228), subjects which were introduced in September 2017, and in which arXiv did not yet have many preprints at the time of data collection.

B. Crossref

We established the publication date using Crossref, which is available in-house at CWTS. We used the Crossref database that was imported on August 2018. We determined the publication date as the first date of the following dates from Crossref: “published online”, “published print”, “created” and “issued. We established the publication date for all arXiv preprints. Out of the 727,186 provided DOIs in arXiv, we find a match in Crossref for 722,003 articles (99%).

We established the publication date for all citing publications using Crossref in the same way. See the next paragraph for more details concerning the citing publications.

C. Scopus

The Scopus database is available in-house at CWTS, which we used for our analysis. We relied on the Scopus database that was imported on May 2018.

We used Scopus to find the published version of the preprint. This was done by matching the DOI from arXiv with the DOI as recorded in Scopus. Out of the 722,003 DOIs from arXiv that were matched to Crossref, we found 664,741 DOIs from Scopus with a unique match (92%). We used the matched publication in Scopus to identify the journal in which the preprint was published.

We calculated the impact of journals using Scopus. We defined the impact as the average number of citations received in the first five years after publication for all articles (document type ar) and reviews (document type re).

a) Electronic mail: v.a.traag@cwts.leidenuniv.nl
For articles that were published within five years of the end of the database (2018), we counted citations until the end of the database.

Finally, we used Scopus to identify citations of both the preprint version and the published version. We parsed all raw cited reference strings provided in Scopus to extract an arXiv identifier or a DOI. We identified arXiv identifiers in the reference string using the regular expression

\[
[a-zA-Z-2\-\. ] + / ?[0-9]{7,} / [a-zA-Z] [rR] [xX] [iI] [vV]:
[0-1][0-9](\{0\} [0-9] [1][0-2]) [0-9] \{4,5\}
\]

If the reference was matched by Scopus, and a cited publication was identified, we used the DOI from the cited publication as recorded in Scopus. If that was not available, we used the DOI in the reference string extracted using the regular expression

\[b10\. [0-9] \{4,\} (\./ [0-9]+) */ S* \b\]

We identified 4,679,896 references with arXiv identifiers in more than half a billion references in total.

For all citing documents, we extracted the publication date through Crossref, as described earlier. We used this date as the cited date of the cited document. The cited date is used at the resolution of a day. Note that some reference may still cite the preprint, even if the preprint is published, although most citations after the preprint is published refer to the published version, as already observed earlier by Larivière et al. (2014). For clarity, we define citations that were made on or before the publication date of the preprint as pre-publication citations, and we define citations that were made after the publication date as post-publication citations. In total we identified 156,528 pre-publication citations and 15,939,887 post-publication citations from references in Scopus.

II. MODEL AND BAYESIAN INFERENCE

There is a clear selection bias (Bareinboim and Pearl, 2012) on papers being submitted to the arXiv or not \(A\). We assume that the latent citation rate \(\phi\) may affect whether a paper will be submitted to the arXiv \(A\), which in turn may affect the journal \(J\). Previous research showed that publications that are available as preprints are more highly cited (Larivière et al., 2014; Fu and Hughey, 2019), but this “citation advantage” seemed unlikely to be causal (Davis et al., 2008; Gaulé and Maystre, 2011). We therefore assume the arXiv does not directly influence the citations \(C\). If we control for \(\phi\) (which is effectively done by controlling for \(C'\)), we obtain that \(\Pr(C \mid do(J), A = 1, \phi) = \Pr(C \mid do(J), \phi)\) by the rules of do-calculus (Pearl, 2009). We thus obtain an unbiased estimate of the causal effect \(\Pr(C \mid do(J))\), even if our observations are biased towards arXiv papers, as stated in the main text.

As explained previously, the journal causal effect \(\Pr(C \mid do(J))\) is not affected by the selection on arXiv papers \(A\). The same does not hold for the estimate \(\Pr(J \mid do(\phi))\), as \(A\) could possibly act as a mediator. Possibly, authors decide to to only post preprints they deem sufficiently good. Being posted on the arXiv may possibly affect where it is published, for example, because some journals may have policies against publishing preprints. In our causal diagram, \(\phi\) may then affect \(A\) which in turn may affect \(J\). Because of the selection effect on \(A\), the effect of \(\phi\) on \(J\) then perhaps only holds for arXiv preprints. To better understand this possible mediating effect, we computed for each journal the proportion of arXiv papers it published. We only included arXiv papers that had at least a preprint duration of at least 30 days. We find there is no discernible relationship between journal impact and the proportion of arXiv papers (Fig. S7). In other words, \(A\) is unlikely to act as a mediator, suggesting that high-impact journals indeed select articles with higher latent citation rates. Although this observation is again confounded by the latent citation rate \(\phi\), it would be rather surprising to have a confounding effect that exactly cancels out the actual causal effect of \(A\) on \(J\), so that we observe no correlation between \(A\) and \(J\).

The full specification of the hierarchical Bayesian model introduced in the main text is as follows. As already introduced in the main text, we model the probability of attracting \(c_i(t)\) citations at time \(t\) as

\[c_i(t) \sim \text{Poisson}(\lambda_i(t)f_i(t) (m + C_i(t - 1))) \tag{1}\]

with \(m\) some parameter affecting the initial rate of attracting citations and

\[\lambda_i(t) = \begin{cases} \phi_i & t \leq T_i' \\ \phi_i \theta_i & t > T_i' \end{cases}, \tag{2}\]

where \(T_i'\) is the date at which publication \(i\) is published in a journal with \(t = 0\) the time at which the preprint was posted on arXiv. We are modelling citations at a daily rate, and it is reasonable to assume that citations on the
same day have not influenced each other. Citations on the same day can be regarded as independent events. The Poisson distribution models exactly a random variable that counts the number of events that happen at a given rate within a given interval, making it a suitable distribution for \( c_i(t) \). This is a slight generalization from the earlier model by Wang, Song, and Barabási (2013) who only consider the probability of being cited at a certain time \( t \). In practice, publications may attract multiple citations at a single day, and we therefore consider the number of citations explicitly. This happens only infrequently, as only about 6% of the days at which a publication is cited is it cited more than once in our dataset.

The temporal decay is represented by \( f_i(t) \), which follows the density of an exponential distribution

\[
f_i(t) = \int_t^{t+1} \frac{1}{\beta} \exp \left[ -\frac{\tau}{\beta} \right] d\tau
\]

\[
= \exp \left[ -\frac{t}{\beta} \right] - \exp \left[ -\frac{t+1}{\beta} \right]
\]  

(3)

(4)

We define \( F_i(t) = \sum_{\tau=0}^{t} f_i(\tau) \), so that

\[
F_i(t) = 1 - \exp \left[ -\frac{t+1}{\beta} \right].
\]  

(5)

For the temporal decay we assume a prior of

\[
\beta_i \sim \text{InvGamma}(2, 3 \times 365).
\]

(6)

Our prior expectation is that the decay takes about 3 years, which corresponds roughly to earlier results (Wang, Song, and Barabási, 2013). This agrees also with other literature on the decay of citations (Egghe and Ravichandra rao, 1992; Avramescu, 1979; Parolo et al., 2015). Note that we do not use the log-normal distribution for the decay, as used in the model on which we build (Wang, Song, and Barabási, 2013). Modelling the decay using the log-normal distribution resulted in problem of convergence, which seemed to be due to multimodality of the logarithmic decay, problematizing model identifiability. Using a maximum likelihood approach as used in the earlier work (Wang, Song, and Barabási, 2013) may miss this multimodality. Using an exponential decay improved the convergence of the Bayesian sampling. Note that even an exponential decay can lead to an initial increase of the number of citations and later decrease, as is typical of citations. We show this in when analyzing the model in more detail in the next section.

There is a certain degeneracy in the model for pre-publication citations that depends on our assumptions of the prior for the decay. If we observe few pre-publication citations, this can be due to two factors: a low decay \( f_i(t) \) at that point \( t \), or a low \( \phi_i \). It is therefore important to assume reasonable priors for the temporal decay. If we assumed that \( f_i(t) \) would be mostly concentrated in the first few days, we would erroneously infer a too low \( \phi_i \) and a too high \( \theta_j \). Although an exponential decay by definition only decreases, our prior expectation is that the decay is quite gradual. The prior on \( \beta_i \) is also quite broad, allowing for substantially different decay.

We assume that the latent citation rate of articles published in a certain journal \( j \) is distributed as

\[
\phi_i \sim \text{LogNormal}(\Phi_j, \epsilon_j).
\]  

(7)

We assume priors of

\[
\Phi_j \sim \text{Normal}(0, 1),
\]

(8)

\[
\epsilon_j \sim \text{InvGamma}(2, 1).
\]

(9)

which roughly corresponds to distributions of \( \lambda_j \) as found in (Wang, Song, and Barabási, 2013) for various journals, assuming the journal citation multiplier is about 1. Although \( \phi_i \) is modelled hierarchically as an element of a journal, causally speaking, \( \phi \) determines \( J \), not the other way around. That is, there is certain causal effect \( \Pr(J \mid \text{do}(\phi)) \), which we assume to give rise to the probability \( \Pr(\phi \mid J) \) we model here. The use of priors in fitting this type of models is also employed in a response (Wang et al., 2014) to some critique of the model (Wang, Mei, and Hicks, 2014). In line with (Wang et al., 2014) we simply set \( m = 30 \) and do not infer this parameter from the data.

Finally, we assume the following prior on the journal citation multiplier \( \theta_j \)

\[
\theta_j \sim \text{Gamma}(2, 2),
\]

(10)

which is centered around 1.

The larger citation rates observed for high-impact journals may correspond to either a higher \( \Phi_j \) or a higher \( \theta_j \). Our priors are relatively conservative with respect to a journal causal effect. We have assumed a prior on \( \Phi_j \) that
corresponds to overall distribution citation rates as found in earlier work (Wang, Song, and Barabási, 2013). The prior on \( \theta_j \) is centered around 1, corresponding to no journal causal effect, but still allows for larger \( \theta_j \).

We use \texttt{pystan 2.19.0} to perform Bayesian inference of the posterior distributions using the no-U-turn sampler (Stan Development Team, 2017). In practice, citations are relatively sparsely distributed throughout time and \( c_i(t) = 0 \) for most \( t \). Instead of specifying the probability for each \( t \) separately, we can more efficiently specify the probability for only those \( t \) for which \( c_i(t) > 0 \). The probability of observing 0 citations for a duration of \( \tau \) is identical to an exponential distribution with the same rate as the Poisson distribution in Eq. 1. More specifically, for a \( t_1 \) and \( t_2 \) such that \( c_i(t_1) > 0 \) and \( c_i(t_2) > 0 \), the probability of observing 0 citations for all \( t \) between \( t_1 \) and \( t_2 \) then equals

\[
\Pr(C_i(t_2 - 1) - C_i(t_1) = 0) = \exp\left[-\lambda_i(t_1)(m + C_i(t_1))(F_i(t_2 - 1) - F_i(t_1))\right]
\]

(11)

assuming times \( t_1 \) and \( t_2 \) do not cross the publication date \( T'_i \). In they do cross \( T'_i \), the time windows \((t_1, T'_i)\) and \((T'_i, t_2)\) should be considered separately. To improve the numerical stability of \texttt{pystan}, we use a logarithmic specification of the rate for the Poisson distribution. This also necessitates to work with the logarithm of the temporal decay, which has a simple form. Finally, we use four chains of 1 000 iterations each, using half of the iterations for warmup with a target acceptance rate of 0.98 (\texttt{adapt.delta}) and a maximum tree depth of 20.

We perform our analysis per year (2000–2016) and field, and restrict to journals that have at least 20 articles that were published at least 30 days after being posted as a preprint on arXiv (Fig. S1). This results in 3 892 different subsets that are separately fitted. The different subsets cover 258 different journals. There were seven subsets which were published at least 30 days after being posted as a preprint on arXiv (Fig. S1). This results in 3 892 different subsets which yielded diverging transitions. Only one subset showed large problems, and almost 25% of the transitions diverged. Nonetheless, we excluded all subsets that showed diverging transitions, but results are unaffected by the exclusion or inclusion of these seven problematic subsets. Using log-normal temporal decay resulted in diverging transitions for about two-third of the subsets.

Source code for fitting our model is available in the Zenodo repository https://doi.org/10.5281/zenodo.3583012.

A. Analysis

We first analyse the mean number of citations attracted by article \( i \). We can write the total number of citations \( C_i \) as \( C_i(t) = C_i(t - 1) + c_i(t) \) for \( t > 0 \) with \( C_i(0) = c_i(0) \). Taking the expected value then yields

\[
E(C_i(t)) = E(C_i(t - 1)) + E(c_i(t)).
\]

(12)

Writing out the expected number of citations received at time \( t \) yields

\[
E(c_i(t)) = \sum_{C=0}^{\infty} E(c_i(t) \mid C_i(t - 1) = C) \Pr(C_i(t - 1) = C)
= \sum_{C=0}^{\infty} \lambda_i(t)f_i(t)(m + C) \Pr(C_i(t - 1) = C)
= \lambda_i(t)f_i(t)(m + E(C_i(t - 1))),
\]

so that we end up with the recursion

\[
E(C_i(t)) = E(C_i(t - 1)) + \lambda_i(t)f_i(t)(m + E(C_i(t - 1))).
\]

(13)

This recursion has as a solution

\[
E(C_i(t)) = m \left( \prod_{\tau=0}^{t} (1 + \lambda_i(\tau)f_i(\tau)) - 1 \right),
\]

(14)
which can be easily checked by substituting in Eq. (13):

$$E(C_i(t)) = E(C_i(t - 1)) + \lambda_i(t) f_i(t)(m + E(C_i(t - 1)))$$

$$= m \left( \prod_{\tau=0}^{t-1} (1 + \lambda_i(\tau)f_i(\tau)) - 1 \right) + \lambda_i(t) f_i(t) \left( m + m \prod_{\tau=0}^{t-1} (1 + \lambda_i(\tau)f_i(\tau)) - 1 \right)$$

$$= m \left( \prod_{\tau=0}^{t-1} (1 + \lambda_i(\tau)f_i(\tau)) - 1 + \lambda_i(t) f_i(t) \prod_{\tau=0}^{t-1} (1 + \lambda_i(\tau)f_i(\tau)) \right)$$

$$= m \left( \prod_{\tau=0}^{t} (1 + \lambda_i(\tau)f_i(\tau)) - 1 \right).$$

Writing the product as an exponential sum of logarithms we obtain

$$E(C_i(t)) = m \left( \exp \left[ \sum_{\tau=0}^{t} \log(1 + \lambda_i(\tau)f_i(\tau)) \right] - 1 \right). \quad (15)$$

A simple Taylor expansion shows that $\log(1 + x) \approx x$ for small $x$, so that we obtain the approximation

$$E(C_i(t)) \approx m \left( \exp \left[ \sum_{\tau=0}^{t} \lambda_i(\tau)f_i(\tau) \right] - 1 \right). \quad (16)$$

Expanding $\lambda_i(\tau)$ we obtain

$$E(C_i(t)) \approx \begin{cases} 
    m (\exp [\phi_i F_i(T'_i)] + \phi_i \theta_j (F_i(t) - F_i(T'_i))] - 1 & \text{for } t \leq T'_i \\
    m (\exp [\phi_i F_i(T'_i)] - 1) & \text{for } t > T'_i.
\end{cases} \quad (17)$$

The expected number of pre-publication citations is given by $E(C_i) = E(C_i(T'_i))$ while the expected number of post-publication citations is given by $E(C_i) = E(C_i(T_i)) - E(C_i(T'_i))$ so that we obtain respectively

$$E(C'_i) \approx m (\exp [\phi_i F_i(T'_i)] - 1) \quad (18)$$

and,

$$E(C_i) \approx m (\exp [\phi_i F_i(T'_i) + \phi_i \theta_j (F_i(T_i) - F_i(T'_i))] - 1) - m (\exp [\phi_i F_i(T'_i)] - 1)$$

$$= m \exp [\phi_i F_i(T'_i)] (\exp [\phi_i \theta_j (F_i(T_i) - F_i(T'_i))] - 1). \quad (19)$$

Taking the limit of $t \to \infty$ and assuming pre-publication duration is negligible, we obtain the approximation of the expected number of long-term citations of $m(e^{\phi_i \theta_j \lambda_i} - 1)$.

Using the approximation for the total number of citations $E(C_i(t))$ we can also obtain an approximation for the expected instantaneous number of citations. This approximation shows that the number of citations can initially increase, even if the temporal decay is exponential. We use a continuous time approximation, and take the derivative of Eq. 16 with respect to $t$ and assume $\theta = 1$ for simplicity. We then obtain the approximation that

$$E(c_i(t)) \approx \frac{m \phi_i}{\beta} \exp \left[ \phi_i \left( 1 - e^{-t/\tau} \right) - \frac{t}{\beta_i} \right], \quad (21)$$

which attains its maximum at $t = \beta_i \log \phi_i$ for $\phi_i > 1$. This shows that citations first increase and then decrease, similar to what is observed empirically. Publications with a slower decay attain this peak later. Similarly, publications that have a higher latent citation rate also attain the maximum at a later time. Interestingly, this is formally equivalent to an older result (Avramescu, 1979).

We can also analyse the variance of $C_i(t)$ and obtain the recursion

$$\text{Var}(C_i(t)) = \text{Var}(C_i(t - 1)) + \text{Var}(c_i(t)) + 2\text{Cov}(C_i(t - 1), c_i(t)). \quad (22)$$

Since $\text{Cov}(C_i(t - 1), c_i(t)) > 0$ this recursion yields a variance $\text{Var}(C_i(t))$ that is larger than the expected value. Hence, there is considerable uncertainty in citations in this model, even for an exact $\phi_i$ and $\theta_j$. This means that even for specific $\phi_i$ and $\theta_j$, the distribution of citations would be quite skewed. It is therefore possible that skewed citation
distributions within a journal emerge, even if latent citation rates $\phi_i$ are homogeneously distributed (Waltman and Traag, 2020).

This result is mostly due to the rich-get-richer effect, also known as Matthew effect or cumulative advantage, which is frequently argued to explain the high variance and skewness observed in most citation distributions, dating back to early literature in scientometrics (Price, 1976). Without the rich-get-richer effect, citations $C_i(t)$ would simply be Poisson distributed around $\sum_{t=0}^{\tau} \lambda_i(t)f_i(t)m$ according to this model. In that case, citation distributions tend to be less skewed for specific $\phi_i$, so that the skewness in citation distributions may require a more heterogeneous distribution of $\phi_i$. We cannot distinguish between these two alternative possibilities based on our empirical observations. It would be interesting to empirically substantiate the cumulative advantage effect for citations, but this goes beyond the scope of this paper. In line with previous literature, we assume the presence of a cumulative advantage effect in our model.

REFERENCES


FIG. S1. Preprints on arXiv. (a) the number of preprints submitted to arXiv per day; (b) the time before a preprint is published ($T_i'$). The shaded areas indicate what part of the data is used for estimating the journal causal effect.
FIG. S2. Detailed results. This shows the dependency of the citation multiplier \( \theta_j \), the median latent citation rate \( e^{\Phi_j} \) and the \( \epsilon_j \) on journal impact (a-c). The visualization shows the median and the error bars represent the 95% credible interval. This also shows the same results but separated per year (d-f) and field (g-i).
FIG. S3. Correlation dynamics. The correlation is provided for the five largest fields over time, for years that have at least 20 journals present. We use the median estimate for the journal citation multiplier $\theta_j$ and the latent citation rate $\Phi_j$ to calculate correlations. We take the logarithm of the journal impact, and the logarithm of the median journal citation multiplier $\theta_j$ before calculating the correlations. The correlation between the journal impact and the journal citation multiplier $\theta_j$ seems increasing for High Energy Physics and Astrophysics over time (a), while the correlation between journal impact and latent citation rate $\Phi_j$ is decreasing over time (b).
FIG. S4. Overview per field and year. Distribution of median estimates of $\Phi_j$ and $\theta_j$ for (a) field and (b) year. Error bars indicate 95% percentile intervals of median estimates for journals in specified field or year. There is some variation over fields. The multiplier $\theta_j$ seems to be relatively high for Statistics, whereas Quantitative Finance shows a relatively low multiplier. Possibly, statisticians do not regularly follow new preprints on arXiv. There seems to be some trend over the years of increasing journal citation multipliers but the trend is not very clear.
FIG. S5. Predicted citations versus observed citations. The 95% credible interval of the predicted number of citations is roughly between half and twice the median predicted number of citations. This quantifies both the uncertainty of the inferred parameters as well as the uncertainty arising from the citation dynamics themselves. For lower number of citations the credible interval is a bit broader.

FIG. S6. Median effective citation rates and journal impact.
FIG. S7. Percentage of publications that are available as preprints on arXiv. This is limited to only preprints that have been posted on arXiv at least 30 days before publication.