Does the Market of Citations Reward Reproducible Work?

Edward Raff
Booz Allen Hamilton
University of Maryland, Baltimore County
`raff_edward@bah.com`

Abstract

The field of bibliometrics, studying citations and behavior, is critical to the discussion of reproducibility. Citations are one of the primary incentive and reward systems for academic work, and so we desire to know if this incentive rewards reproducible work. Yet to the best of our knowledge, only one work has attempted to look at this combined space, concluding that non-reproducible work is more highly cited. We show that answering this question is more challenging than first proposed, and subtle issues can inhibit a robust conclusion. To make inferences with more robust behavior, we propose a hierarchical Bayesian model that incorporates the citation rate over time, rather than the total number of citations after a fixed amount of time. In doing so we show that, under current evidence the answer is more likely that certain fields of study such as Medicine and Machine Learning (ML) do correlate reproducible works with more citations, but other fields appear to have no relationship. Further, we find that making code available and thoroughly referencing prior works appear to also positively correlate with increased citations. Our code and data can be found at `https://github.com/EdwardRaff/ReproducibleCitations`.

1 Introduction

A reproducibility crisis has been called for many scientific domains, including artificial intelligence and machine learning (Donoho et al., 2009; Baker, 2016; Hutson, 2018; Vul et al., 2008). It is paramount that all disciplines work to remedy this situation and push for reproducible work both as good science, and to mitigate such crises. Such work has begun in various fields with different strategies (Errington et al., 2021; Poldrack, 2019; Collaboration, 2015; Sculley et al., 2015; Gardner et al., 2018), yet the incentive structure around producing reproducible work has received almost no attention. We note that the difference in terminology between reproduction and replicating is long, with conflicting terminology across fields and years (Plesser, 2018), we will use both terms interchangeably as our study focuses exclusively on cases where a different team independently performs the same experiments to obtain the same/similar results.

Citations are the primary reward for academic outputs, and to our knowledge only the work of Serra-Garcia & Gneezy (2021) has ever considered studying the relationship between papers that reproduce and the number of citations received. They used data on replication results from the fields of Psychology (Collaboration, 2015), Economics (Camerer et al., 2016), and Social Sciences (Camerer et al., 2018). Distressingly, they conclude that non-reproducing work is cited more than reproducing works.

Our work revisits this hypothesis and data, and draws a different conclusion. We will show in section 3 that there are methodological issues that prevent a robust conclusion from being formed with the data and approach presented in (Serra-Garcia & Gneezy, 2021). Next, we will propose a Bayesian hierarchical model to alleviate these issues and allow further insight into the citation/replication question by incorporating a model of the citation rate changing over time in section 4. In section 5 we show our model is a significantly better fit to the data, and concludes that citation rate is unrelated or positively correlated with reproduction success, depending on the field being studied. Finally, we will conclude in section 6.

2 RELATED WORK

The study of paper citation has a long and multi-disciplinary history (Lotka, 1926; Shockley, 1957; Price, 1965; 1976; Potter, 1981; Redner, 1998), with many works proposing different power law variants to describe the distribution of citations. Most work that has looked at citations over time are looking at population level changes in citation distributions (Bornmann & Mutz, 2015; Varga, 2019; Wallace et al., 2009). We are aware of only one prior work that looked at the citation rate by year through studying the impact of publication-vs-arXiv (Traag, 2021). This work also modeled citation rates as a Poisson, similar to Serra-Garcia & Gneezy (2021), which we will argue is an inappropriate model for citation count data.

Used by Serra-Garcia & Gneezy (2021) were *negative citations*, a type of citation classification that can provide further insight into behaviors and results. The taxonomy of citation types, their labeling, and prediction (Kunnath et al., 2022) are another lens through which insight may be gained, but is beyond the scope of our study.

Dietz et al. (2007) produced one of the first applications of Bayesian modeling to the study of citation behavior and influences. Our task is different, and so our model bares little resemblance, but the overall strategy we argue is worth further study. Several latent factors exist in bibliometric study to which modern machine learning may yield benefits, and the scale of bibliometric data provides fertile ground to new and technical challenges to advance the field.

3 ISSUES WITH EXISTING MODELING

While the Negative Binomial model has been previously identified to empirically have better performance at citation prediction (Thelwall & Wilson, 2014), the Poisson model is still very popular. We note though that there is an easier way to show the Poisson model is in fact, inappropriate, for the bibliometric research it is used. The Poisson model assumes the mean and variance are equal, and if the variance is larger than the mean, the model suffers from overdispersion that prevents meaningful results. A statistical test (Cameron & Trivedi, 1990) confirms with $p < 0.001$ that this is the case for citation data, which in the data from (Serra-Garcia & Gneezy, 2021) has a mean of 438 citations but a variance of 504,639.

While Serra-Garcia & Gneezy (2021) used the Poisson model in their work on the connection between replication and reproducibility, we note there are additional factors that lead us to challenge their initial conclusion. The first is a data issue of reproducibility itself: $N = 80$ documents were noted in (Serra-Garcia & Gneezy, 2021), but the data provide $N = 139$ instances. We are unable to determine the correct selection criteria[1] to render only 80, and so proceed forward with the larger number of samples.

Table 1: Results indicating if successfully reproduced papers have more (positive) or less (negative) citations than papers that failed to reproduce. Models tested include Poisson verse NegativeBinomial (NB) regressions using the original three domains with Google Scholar (GS) or Semantic Scholar (SC) citations each, and an additional case using SC with a fourth set of reproduction results from the Medical domain (+M).

	Poisson-GS		Poisson-SC		Poisson-SC+M		NB-GS		NB-SC		NB-SC+M	
	coef	p	coef	p	coef	p	coef	p	coef	p	coef	p
Reproduced	0.0172	0.129	0.1138	<0.001	0.5775	<0.001	0.0172	0.150	4.4592	<0.001	0.5777	0.004

To demonstrate the lack of robustness to the prior methodology, we will perform several repetitions of the overall approach choosing between:

    1. Using the Poisson model versus a Negative-Binomial model

---

[1]The authors graciously spent considerable time working with us, and we did not have the same software licenses to use their saved results. One hypothesis from the authors was that non-significant results were excluded, but only removed 16 samples when we went through the data provided. Cross-discipline reproducibility and data sharing standards poses an interesting question beyond our scope.

2. Using the original Google Scholar (GS) citation count data provided vs citation data from Semantic Scholar (SC)

3. Using the original data with (SC) additionally with reproduction results from the Medical domain, adding a fourth field (+M).

This provides six total results, presented in Table 1 using (Seabold & Perktold, 2010). We can see in that no case do we observe a negative indication that papers which fail to replicate are cited more. However, we do see inconsistent conclusion about the impact of replication itself. When using Google Scholar the conclusion is there is no relationship, and when using Semantic Scholar the conclusion is a strong relationship. This challenge is not a factor of these citations sources having dramatic disagreement, as can be seen in Figure 1 both are highly correlated in the per-year citations of the documents. This issue is instead that of model fit, as the highest adjusted R^2 fit amongst the Negative Binomial models is 0.0039.

The source of this discrepancy is inappropriate merging of all data sources into one pool. The papers selected from Economics, Psychology, Social Science, and Medicine where all selected with biases toward higher citation rates — largely through selection of high impact factor sources. The citation rate per field, or journal, are not the same, as shown in Figure 2. Imbalances in the number of papers from each source that happened to replicate or not amplify spurious noise, resulting in low model fit and unstable conclusions.
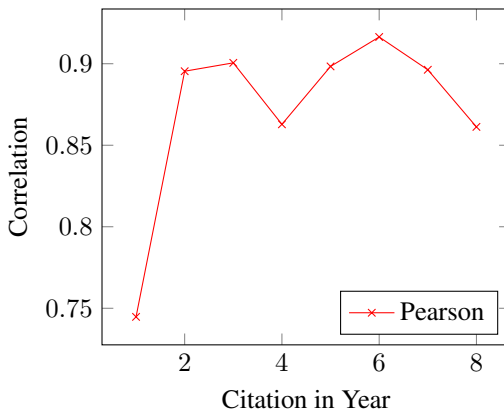

Figure 1: Correlation between Google Scholar and Semantic Scholar in the number of citations for each document per year. After multiple-test correction all years were significantly correlated with $p < 0.001$ in all cases.
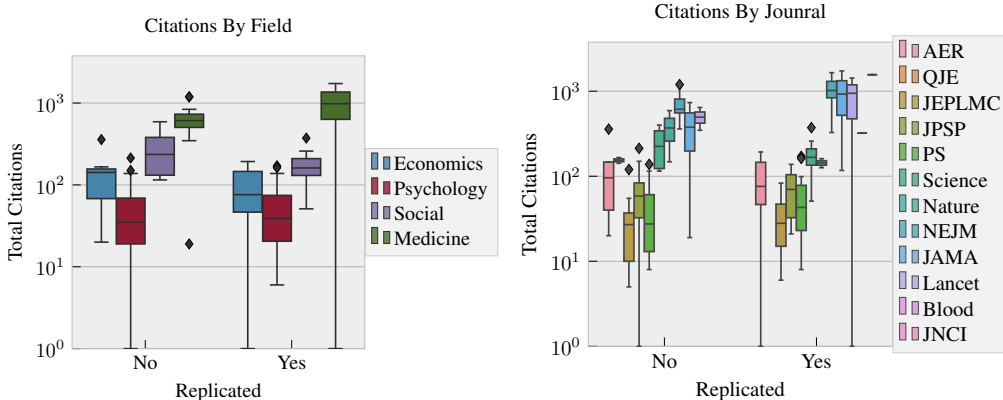

Figure 2: Total number of citations accumulated for replicated and failed to replicate papers grouped by field (left) and journal (right).

4 METHODOLOGY

To address these problems, we propose a Bayesian hierarchical model that incorporates the citation rate over time, rather than the cumulative total number of citations. Our interest in citation rate over time is of interest not merely for model fit, but primarily because we are interested if the types of citation patterns vary between reproducible and non-reproducible papers. That is to say, some papers do not start to accumulate citations for a considerable amount of time, others reach a steady-state of

citations, and others reach a peak citation rate before their citation rate drops. A total-citation rate model can not reveal anything about this question.
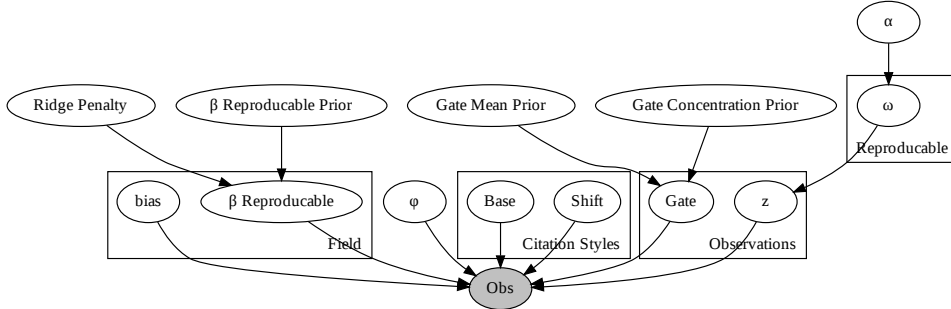

Figure 3: Plate diagram of our proposed citation-replicated model. The observations are done against a Negative-Binomial model.

The high level plate diagram of our approach is presented in Figure 3, which we will discuss at a high level with the detailed generative story given by Algorithm 1. The coefficients β are with respect to each Field, with a hierarchical prior used over then and a shared ridge regression penalty (variance of the Gaussian distribution).

$$\text{NegBinomial2}(n \mid \mu, \varphi) = \binom{n + \varphi - 1}{n} \left(\frac{\mu}{\mu + \varphi}\right)^n \left(\frac{\varphi}{\mu + \varphi}\right)^\varphi. \tag{1}$$

The observations are done with respect to a zero-inflated Negative-Binomial model, parameterized with a mean and dispersion factor μ and φ as shown in Equation 1. The zero-inflation serves two purposes. First, some papers do receive zero citations for some time before becoming popular, and the zero-inflation model prevents down-weighting the citation rate μ from these zero citations. Second, it allows us a convenient way to handle the fact that papers were published at different times, and thus for a desired horizon of T years not all papers will have T years of existence to accumulate citations. When a year has not yet occurred, we force the zero inflation gate to effectively mask the year with no impact on the model. We used a target of $T = 10$ years in all cases. Each paper receives it's own gate value with a hyper prior shared over all samples. We use the proportional Beta hyper prior as shown in Equation 2 with a non-informative prior over μ.

$$\text{BetaProportion}(\theta \mid \mu, \kappa) = \frac{1}{\text{B}(\mu\kappa, (1 - \mu)\kappa)} \theta^{\mu\kappa - 1}(1 - \theta)^{(1-\mu)\kappa - 1} \tag{2}$$

To represent the impact of the t'th year's citation rate of the i'th sample $\mu_{i,t}$ we model a base citation rate μ_i modulated by an annual *base* citation multiplier sampled from a Gamma prior centered at a mean of 1.0 (i.e., no change in annual citation rate). The impact of the compounding base rate can be delayed (but not increased, as that implies pre-publication citations) by a *shift* factor samples from a positive Laplacian scaled so that the entire T years may be selected by the prior would prefer no shift.

We do not give each sample it's own base and shift as it allows significant over-fitting of the model to ignore the impact of the coefficients β. Instead we use a Dirichlet process to sample from a pool base/shift pairs — where reproducible and non-reproducible papers each receive a separate Dirichlet process sampling from the same pool. We enforce a sparse process by putting a Beta prior over the α parameter of the processes so that we may see if there is a difference in the types of citation styles between papers (e.g., do non-replicating papers more frequently have decaying base rates < 1). In each experiment there is one pool of base/shift pairs, and two sets of distributions ω over those pools. One ω_S for reproduced papers and one ω_F for the non-reproduced. In this way the model can

inform us if there appears to be a difference ($\omega_S \neq \omega_F$) in citation styles (base/shift pairs) between the populations.

Algorithm 1 Our Hierarchical Bayesian generative story for modeling citation rates. The $^+$ indicates distributions truncated to be non-negative.

Require: N observations with $r_i \in \{S, F\}$ for successful or failed reproduction and f_i indicating the field of research for the paper.

$\lambda_{ridge} \sim \text{HalfCauchy}(0,1)$

$\alpha \sim \text{Beta}(1, 10)$ *▷A Beta distribution used to encourage sparse solutions*

$\omega^S \sim \text{Dirichlet}(\alpha)$ *▷A different distribution over all base/shift values for reproducible . . .*

$\omega^F \sim \text{Dirichlet}(\alpha)$ *▷and non-reproducible papers*

for all $i \in 1, \ldots, \infty$ **do**   *▷Citation Styles for ω^* will sample from*

     $shift \sim \text{Laplace}^+(0, \text{years out}/6)$

     $base \sim \Gamma(100, 100)$ *▷This Gamma distribution will encourage values near 1, as values ¿ 2 are undesirable in being unrealistic.*

end for

$\widehat{\beta^{field}} \sim \mathcal{N}(0, 1)$ *▷Hierarchical Reproducible Prior*

for all Field of Study i **do**

     $\beta_i^{field} \sim \mathcal{N}(\widehat{\beta^{field}}, \lambda_{ridge})$

     $b_i \sim \text{Cauchy}(0, 1)$ *▷Bias term is independent between Fields*

end for

$\widehat{gate^\mu} \sim \mathcal{U}(0, 1)$ *▷Uninformative prior on the mean rate of no citations occurring.*

$\widehat{gate^\kappa} \sim \Gamma(1, 20)$

$\varphi \sim \text{Cauchy}^+(0, 5)$

for all Observations i **do**

     $z \sim \text{Categorical}(\omega^{r_i})$ *▷Select the citation style base/shift for this sample based on the distribution w.r.t. the sample replicating or not*

     $\log(\mu_i) \leftarrow \beta_{f_i}^{field} \cdot \mathbb{1}[r_i = S] + b_{f_i}$ *▷The rate is modified based on the paper replicating or not.*

     $gate_i \sim \text{BetaProportion}(\widehat{gate^\mu}, \widehat{gate^\kappa})$

     for all Time steps t **do**

        $\mu_{i,t} \leftarrow \mu_i \cdot base_z^{\max(t - shift_z, 0)}$

        accumulate Zero-Inflated Negative Binomial loss $\text{NetBinomial2}(y_i|\mu_{i,t}, \varphi)$ with gate probability $gate_i$

     end for

end for

The full model is detailed in Algorithm 1. We use NumPyro (Phan et al., 2019) to implement the model with the NUTS sampler (Hoffman & Gelman, 2014). In all cases we use 500 burn-in iterations followed by 2,250 steps with a thinning factor of 3.

5 RESULTS

Now that we have specified our approach to understanding how citations may be impacted by a paper's ability to replicate, we will present out results in two sections. First we will consider the results with respect to the previous fields of study (Medicine, Economics, Psychology, Social) and show that we obtain consistent results and reasonably believe them to be a more reliable model. Second we will repeat the study applied to data from machine learning (Raff, 2019). This data is studied separately because it has a different kind of selection bias, and a different set of available features to consider, than the other data.

5.1 SCIENCE RESULTS

We begin by examining the conclusions inferred by our model on the three versions of the data, Google Scholar, Semantic Scholar, and Semantic Scholar with the medical domain added. The results can be found in Figure 4, showing consistent conclusions of no correlation between field and

citation rate of reproducible papers for any of the three original fields. When Medicine is added we observe that it does show high citation rate for reproducing papers, without changing the conclusion of the other fields.
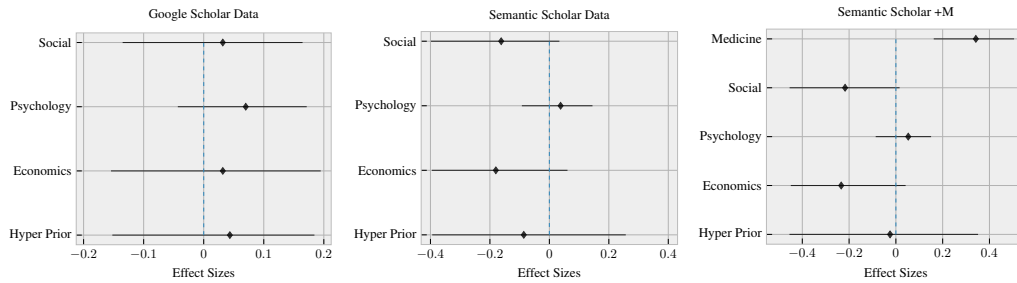

Figure 4: The results of the coefficients β for the different Fields of study when using Google Scholar data (left), Semantic Scholar (middle), and Semantic Scholar with the addition of the medical papers (right). The x-axis is coefficient value and the forest plot shows the estimated value and 95% credible interval.

Beyond the consistency of the conclusions, we are further confident in our approach's conclusions due to better model fit. The Google Scholar case producing an $R^2 = 0.41$, and the Semantic Scholar data with/without the Medicine papers at $R^2 = 0.24$ and 0.19 respectively. We arguably would not expect very high R^2 values considering the model is characterizing populations of citation rates based only on the field, as prior work focusing on predicting citations using venue, author, and content information achieved $R^2 = 0.74$ (Yan et al., 2011).

This approach has also provided further insight into the nature of reproduction and citations, that the reward behaviors are not consistent across fields (subject to unobserved confounders). The question then becomes: do reproducible papers have a different *style* of citation patterns (i.e., accumulating or decreasing citation rates at a different pace) compared to no reproducible work?
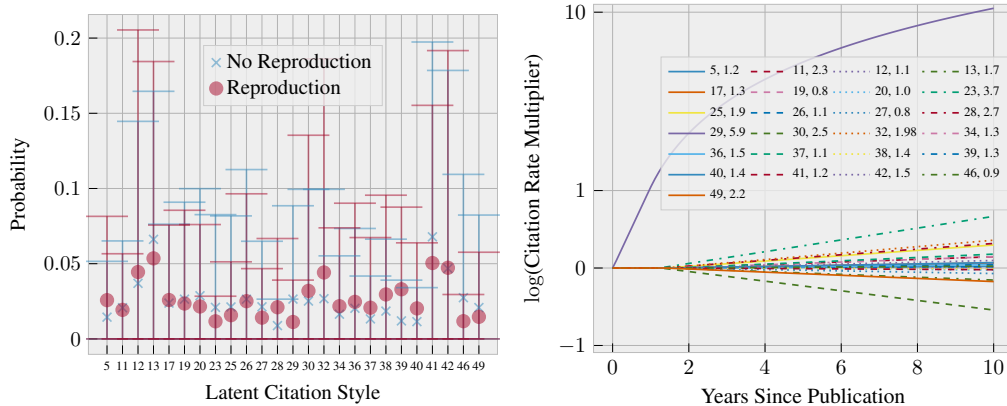

Figure 5: The discovered latent citation styles and their proportion of use in reproduced and failed to reproduce papers (left) and the log multiplicative effect of the citation rate over time (right). Note the right legend shows "Citation Style, Mean Occurrence Rate of the Style". The y-axis is a symmetric logarithm scale with linear behavior in the $[-1, 1]$ range, where 0 indicates no change in the citation rate. Error bars are the 95% credible interval inferred by the model.

Per the design of our model, in Figure 5 we can investigate the citation rates over time as inferred by our model, shown for the Semantic Scholar + Medicine case. In this instance we do not observe any difference in the citation rates or style between (non)reproducing papers. A maximum of 50 components were allowed for computational tractability, and non-present components are ones the model learned to discard with near-zero probabilities. We note of particular interest that most latent citation styles only have an impact starting two years out from publication, a result consistent with prior work which found the first two years of citations to be highly predictive of the long-term

cumulative number of citations (Stegehuis et al., 2015). This provides another degree of confidence in the validity of our general approach, though we do not make claim that our simple model of citation rate is the best possible choice.

The data is also interesting in that we observe behaviors not normally discussed in bibliometric literature: papers who's citation rate decreases with time. This is indeed not directly observable in the common modeling approach of looking at cumulative citations after a point in time. We further find citation style 29 uniquely interesting as a "runaway success", quickly multiplying the citation rate by $\exp(10) \approx 10^{4.35}$ after ten years.

5.2 MACHINE LEARNING RESULTS

Having shown our model allows for more robust conclusions around the impact replicatiable results has on citation rate, we turn to the machine learning reproductions documented by Raff (2019). Many of the papers were selected by the author's personal interests, rather than impact factor, so we do not find it appropriate to include it in the same hierarchical model. The ML data also includes numerous other quantification's about the paper not present in the prior section, so we treat it separately. We use the same approach without a hierarchical prior over field since it is one population of papers. The adjusted R^2 of the model is 0.31 using Semantic Scholar for the citation data, inline with the prior experiments.
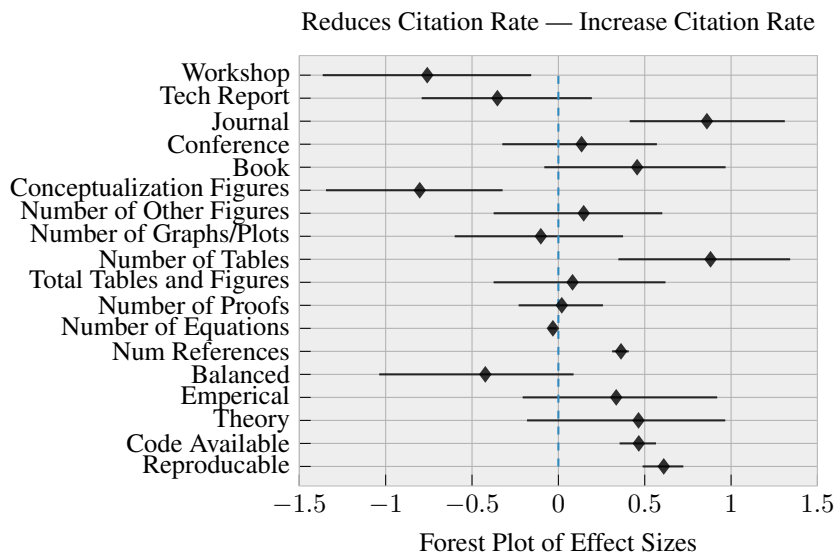

Figure 6: Forest plot of the coefficients β of various features, with 95% credible intervals.

The results Figure 6 show that reproducible papers, and papers that make their code available, both receive higher rates of citation. The former is desirable, and the later indicates a strong motivation for authors to open source their code beyond the arguments around replication (Kluyver et al., 2016; Claerbout & Karrenbach, 1992; Callahan et al., 2016; Errington et al., 2021; Forde et al., 2018). The sharing of code is generally argued to be beneficial, but we do note that it captures methodological flaws as well — and is thus not a panacea to concerns around reproduction (Dror et al., 2019; 2017; 2018; Sun et al., 2020; Bouthillier et al., 2019; 2021). We are also encouraged that more references per page has a higher citation rate, under the belief this corresponds to more thorough documentation of prior work and good scholastic behaviors.

The reduced citation rate for Conceptualization Figures, which attempt to convey the intuition of a method, is interesting. Raff (2019) noted no relationship between this variable and replication, while later work found that papers which use conceptualization figures take less time/human effort to reproduce (Raff, 2021). This type of scientific communication appears to have a particularly complex relationship with reproduction and the incentives around reproduction that thus warrants further study.

The last points of note are that publishing in Journals, and more tables appear to increase citation rate while publishing in a workshop reduces it. Publishing in a workshop having a lower citation rate makes sense intuitively, though it is perhaps interesting that tech reports (like arXiv) have no relationship — and it is worth studying whether workshops being a final "home" for a paper may have a negative perception. This result is also possible due to the noted bias in the data, which we believe may explain the result that Journal publications have a higher citation rate, since ML as a field generally prefers conferences over journals. Last, we have no particular intuition about why having more tables per paper may lead to more citations — unless it is a matter of making it easy for future papers to re-use the reported results, a hypothesis proposed in (Raff, 2019).
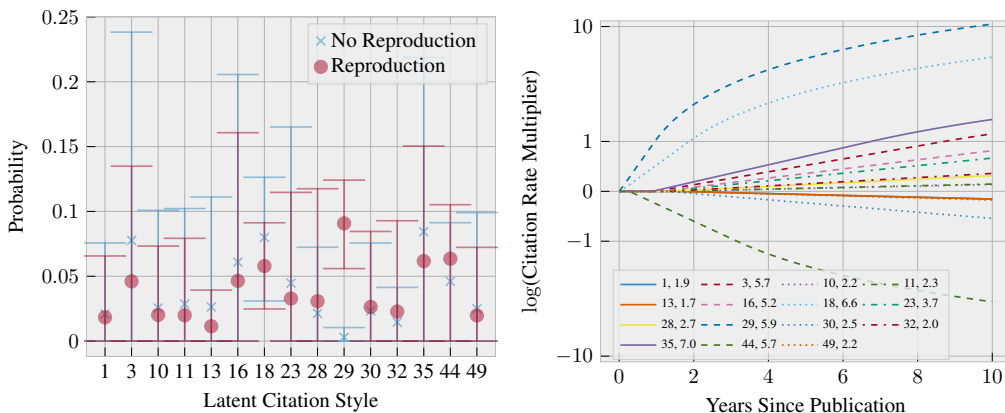

Figure 7: The discovered latent citation styles and their proportion of use in reproduced and failed to reproduce papers (left) and the log multiplicative effect of the citation rate over time (right) for the Machine Learning data. Note the right legend shows "Citation Style, Mean Occurrence Rate of the Style". The y-axis is a symmetric logarithm scale with linear behavior in the $[-1, 1]$ range, where 0 indicates no change in the citation rate.

Last, we look at the latent citation patterns again in Figure 7, and note that style 29 does has a significant difference between reproducible and non-reproducible works[2]. By chance this component again represents a "runaway success", indicating a preference for degree of success toward reproducible works.

6 CONCLUSION

Our results are overall encouraging toward the question of replication and citation: citations are positively correlated or are independent of replication, which is better than the prior hypothesis that non-reproducible works get more citations. Our results for machine learning in particular indicate that citations correlate positively with further desirable behaviors like thorough citations and sharing of code. This work has furthered the bibliometric study of the interaction between citation rate and replication, and we note further valuable directions remain. A large amount of data without ground-truth replication success exists to merge into such analyses, as well as the possibility of using natural language processing to make inferences about paper replications by the content of citing documents.

REFERENCES

Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454, 2016. ISSN 1476-4687. doi: 10.1038/533452a. URL https://doi.org/10.1038/533452a.

Lutz Bornmann and Rüdiger Mutz. 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, 2015. ISSN 23301643. doi: 10.1002/asi.23329.

---

[2]This is a different style "29" than in the prior section, it is by pure chance that it also happened to be the 29'th component of the same nature. This took us many hours to "debug" to find no apparent but, just random chance.

Xavier Bouthillier, César Laurent, and Pascal Vincent. Unreproducible Research is Reproducible. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 725–734, Long Beach, California, USA, 2019. PMLR. URL `http://proceedings.mlr.press/v97/bouthillier19a.html`.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Naz Sepah, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Dmitriy Serdyuk, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for Variance in Machine Learning Benchmarks. In *Machine Learning and Systems (MLSys)*, 2021. URL `http://arxiv.org/abs/2103.03098`.

Benjamin Callahan, Diana Proctor, David Relman, Julia Fukuyama, and Susan Holmes. REPRODUCIBLE RESEARCH WORKFLOW IN R FOR THE ANALYSIS OF PERSONALIZED HUMAN MICROBIOME DATA. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 21:183–94, 2016. ISSN 2335-6936. URL `http://www.ncbi.nlm.nih.gov/pubmed/26776185http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4873301`.

Colin F. Camerer, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu. Evaluating replicability of laboratory experiments in economics. *Science*, 351 (6280):1433–1436, 3 2016. ISSN 0036-8075. doi: 10.1126/science.aaf0918. URL `https://www.science.org/doi/10.1126/science.aaf0918`.

Colin F. Camerer, Anna Dreber, Felix Holzmeister, Teck Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A. Nosek, Thomas Pfeiffer, Adam Altmejd, Nick Buttrick, Taizan Chan, Yiling Chen, Eskil Forsell, Anup Gampa, Emma Heikensten, Lily Hummer, Taisuke Imai, Siri Isaksson, Dylan Manfredi, Julia Rose, Eric Jan Wagenmakers, and Hang Wu. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–644, 2018. ISSN 23973374. doi: 10.1038/s41562-018-0399-z.

A.Colin Cameron and Pravin K Trivedi. Regression-based tests for overdispersion in the Poisson model. *Journal of Econometrics*, 46(3):347–364, 1990. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(90)90014-K. URL `https://www.sciencedirect.com/science/article/pii/030440769090014K`.

Jon F. Claerbout and Martin Karrenbach. Electronic documents give reproducible research a new meaning. In *SEG Technical Program Expanded Abstracts 1992*, pp. 601–604. Society of Exploration Geophysicists, 1 1992. doi: 10.1190/1.1822162. URL `http://library.seg.org/doi/abs/10.1190/1.1822162`.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349 (6251), 2015. ISSN 0036-8075. doi: 10.1126/science.aac4716. URL `https://science.sciencemag.org/content/349/6251/aac4716`.

Laura Dietz, Steffen Bickel, and Tobias Scheffer. Unsupervised Prediction of Citation Influences. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pp. 233–240, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595937933. doi: 10.1145/1273496.1273526. URL `https://doi.org/10.1145/1273496.1273526`.

David L. Donoho, Arian Maleki, Inam Ur Rahman, Morteza Shahram, and Victoria Stodden. Reproducible Research in Computational Harmonic Analysis. *Computing in Science & Engineering*, 11(1):8–18, 1 2009. ISSN 1521-9615. doi: 10.1109/MCSE.2009.15. URL `http://ieeexplore.ieee.org/document/4720218/`.

Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. Replicability Analysis for Natural Language Processing: Testing Significance with Multiple Datasets. *Transactions of the Association for Computational Linguistics*, 5:471–486, 11 2017. ISSN 2307-387X. doi: 10.1162/tacl{_}a{_}00074. URL `https://doi.org/10.1162/tacl_a_00074`.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1383–1392, Melbourne, Australia, 7 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1128. URL https://aclanthology.org/P18-1128.

Rotem Dror, Segev Shlomov, and Roi Reichart. Deep Dominance - How to Properly Compare Deep Neural Models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2773–2785, Florence, Italy, 7 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1266. URL https://aclanthology.org/P19-1266.

Timothy M Errington, Alexandria Denis, Nicole Perfito, Elizabeth Iorns, and Brian A Nosek. Reproducibility in Cancer Biology: Challenges for assessing replicability in preclinical cancer biology. *eLife*, 10:e67995, 12 2021. ISSN 2050-084X. doi: 10.7554/eLife.67995. URL https://doi.org/10.7554/eLife.67995.

Jessica Forde, Tim Head, Chris Holdgraf, Yuvi Panda, Fernando Perez, Gladys Nalvarte, Benjamin Ragan-kelley, and Erik Sundell. Reproducible Research Environments with repo2docker. In *Reproducibility in ML Workshop, ICML'18*, 2018.

Josh Gardner, Christopher Brooks, and Ryan S Baker. Enabling End-To-End Machine Learning Replicability : A Case Study in Educational Data Mining. In *Reproducibility in ML Workshop, ICML'18*, 2018.

Matthew D Hoffman and Andrew Gelman. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(47):1593–1623, 2014. URL http://jmlr.org/papers/v15/hoffman14a.html.

Matthew Hutson. Artificial intelligence faces reproducibility crisis. *Science*, 359(6377):725–726, 2018. ISSN 0036-8075. doi: 10.1126/science.359.6377.725. URL https://science.sciencemag.org/content/359/6377/725.

Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, Carol Willing, and Jupyter development team. Jupyter Notebooks - a publishing format for reproducible computational workflows. In Fernando Loizides and Birgit Scmidt (eds.), *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pp. 87–90. IOS Press, 2016. URL https://eprints.soton.ac.uk/403913/.

Suchetha N Kunnath, Drahomira Herrmannova, David Pride, and Petr Knoth. A meta-analysis of semantic classification of citations. *Quantitative Science Studies*, 2(4):1170–1215, 2 2022. ISSN 2641-3337. doi: 10.1162/qss{_}a{_}00159. URL https://doi.org/10.1162/qss_a_00159.

Alfred J Lotka. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16(12):317–323, 10 1926. ISSN 00430439. URL http://www.jstor.org/stable/24529203.

Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro. *arXiv*, pp. 1–10, 2019. URL http://arxiv.org/abs/1912.11554.

Hans E Plesser. Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Frontiers in neuroinformatics*, 11:76, 1 2018. ISSN 1662-5196. doi: 10.3389/fninf.2017.00076. URL https://pubmed.ncbi.nlm.nih.gov/29403370https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5778115/.

Russell A. Poldrack. The Costs of Reproducibility. *Neuron*, 101(1):11–14, 1 2019. ISSN 08966273. doi: 10.1016/j.neuron.2018.11.030. URL https://linkinghub.elsevier.com/retrieve/pii/S0896627318310390.

William Gray Potter. Lotka's Law Revisited. *Library Trends*, 30:21–39, 1981.

Derek De Solla Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306, 1976. doi: https://doi. org/10.1002/asi.4630270505. URL https://asistdl.onlinelibrary.wiley.com/ doi/abs/10.1002/asi.4630270505.

Derek J. de Solla Price. Networks of Scientific Papers. *Science*, 149(3683):510–515, 7 1965. ISSN 0036-8075. doi: 10.1126/science.149.3683.510. URL https://www.science.org/doi/ 10.1126/science.149.3683.510.

Edward Raff. A Step Toward Quantifying Independently Reproducible Machine Learning Research. In *NeurIPS*, 2019. URL http://arxiv.org/abs/1909.06674.

Edward Raff. Research Reproducibility as a Survival Analysis. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence*, 2021. URL http://arxiv.org/abs/2012.09932.

S Redner. How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B - Condensed Matter and Complex Systems*, 4(2):131–134, 1998. ISSN 1434-6036. doi: 10.1007/s100510050359. URL https://doi.org/10.1007/ s100510050359.

D Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. Hidden Technical Debt in Machine Learning Systems. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pp. 2503–2511, Cambridge, MA, USA, 2015. MIT Press.

Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

Marta Serra-Garcia and Uri Gneezy. Nonreplicable publications are cited more than replicable ones. *Science Advances*, 7(21):eabd1705, 5 2021. ISSN 2375-2548. doi: 10. 1126/sciadv.abd1705. URL https://advances.sciencemag.org/lookup/doi/10. 1126/sciadv.abd1705.

William Shockley. On the Statistics of Individual Variations of Productivity in Research Laboratories. *Proceedings of the IRE*, 45(3):279–290, 1957. ISSN 0096-8390. doi: 10.1109/JRPROC. 1957.278364. URL http://ieeexplore.ieee.org/document/4056505/.

Clara Stegehuis, Nelly Litvak, and Ludo Waltman. Predicting the long-term citation impact of recent publications. *Journal of Informetrics*, 9(3):642–657, 2015. ISSN 18755879. doi: 10.1016/j.joi. 2015.06.005.

Zhu Sun, Di Yu, Hui Fang, Jie Yang, Xinghua Qu, Jie Zhang, and Cong Geng. Are We Evaluating Rigorously? Benchmarking Recommendation for Reproducible Evaluation and Fair Comparison. In *Fourteenth ACM Conference on Recommender Systems*, RecSys '20, pp. 23–32, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375832. doi: 10.1145/ 3383313.3412489. URL https://doi.org/10.1145/3383313.3412489.

Mike Thelwall and Paul Wilson. Regression for citation data: An evaluation of different methods. *Journal of Informetrics*, 8(4):963–971, 10 2014. ISSN 17511577. doi: 10.1016/j.joi.2014.09.011. URL https://linkinghub.elsevier.com/retrieve/ pii/S1751157714000923.

V A Traag. Inferring the causal effect of journals on citations. *Quantitative Science Studies*, 2(2): 496–504, 7 2021. ISSN 2641-3337. doi: 10.1162/qss{_}a{_}00128. URL https://doi. org/10.1162/qss_a_00128.

Attila Varga. Shorter distances between papers over time are due to more cross-field references and increased citation rate to higher-impact papers. *Proceedings of the National Academy of Sciences of the United States of America*, 116(44):22094–22099, 2019. ISSN 10916490. doi: 10.1073/pnas.1905819116.

Edward Vul, Christine Harris, Piotr Winkielman, and Harold Pashler. Voodoo Correlations in Social Neuroscience. *Perspectives on Psychological Science*, 2008.

Matthew L. Wallace, Vincent Larivière, and Yves Gingras. Modeling a century of citation distributions. *Journal of Informetrics*, 3(4):296–303, 2009. ISSN 17511577. doi: 10.1016/j.joi.2009.03.010.

Rui Yan, Jie Tang, Xiaobing Liu, Dongdong Shan, and Xiaoming Li. Citation Count Prediction: Learning to Estimate Future Citations for Literature. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pp. 1247–1252, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450307178. doi: 10.1145/2063576.2063757. URL https://doi.org/10.1145/2063576.2063757.

A   APPENDIX

We note that we have released a modified version of the ML citation data at the URL specified. In initial release of the data we kept paper titles withheld due to concern that marking nearly 100 papers as "not reproducible by this author in their attempts" would be misconstrued as "not reproducible", cause potentially high stress for junior authors in the list, and meet with potential acrimony at large. While we believe time may have cooled some concerns, the large number of initial emails, sometimes heated, about the work makes us fear that such self-censorship was unfortunately the best choice of action. As such we are striking a balance that in releasing a version of the data with citations by year, would be too trivially easy to determine the entire author list.

As such a small amount of noise has been added such that the results are generally identical in re-running the analysis, but keeps the reverse of the names at least not completely trivial. Our hope was to use differential privacy to perform a more robust release, but the nature of citation count data meant that the amount of required noise had a hugely detrimental impact on the results that prevented any replication.

We hope the reader will understand that balance we are trying to make, and that the data is still useful.