

# The role of open code in scholarly Communication: A citation analysis in computational linguistics

Jianguen He<sup>1</sup>, Changyang Feng<sup>2</sup>, Wen Lou<sup>3</sup>, Bo Song<sup>4</sup>, Yizhou Zang<sup>5</sup>

<sup>1</sup>*jiangen@utk.edu*

School of Information Sciences, The University of Tennessee, Knoxville, Tennessee, (USA)

<sup>2</sup>*cyfeng@mail.ccnu.edu.cn*

School of Information Management, Central China Normal University, Wuhan, (China)

<sup>3</sup>*wlou@infor.ecnu.edu.cn*

Department of Information Management, Faculty of Economics and Management, East China Normal University, Shanghai, (China)

<sup>4</sup>*bosong@drexel.edu*

College of Computing & Informatics, Drexel University, Philadelphia, Pennsylvania, (United States)

<sup>5</sup>*yizhouzang@gmail.com*

Philips Lighting, Boston, Massachusetts, (United States)

## Abstract

This paper investigated 12 years of full-text papers published in two conferences of computational linguistics. We investigated open code behaviors among researchers and gain insights into the effects of open code on citations. We found that the percentage of open-code papers increased by nearly one-third between 2006 and 2017. The papers with open code have more citations than papers without open code on average for most of the years and our regression models suggested that open code has a significant predictive effect on citations. Our results show that open code can be an effective way to receive visibility and attention for scholarly communication.

## Introduction

Open code is becoming a prominent way of scholarly communication, which has influenced scientific conversation, thought and behavior (Haustein 2016; Piwowar 2013). Open code is a process in which code, models, and algorithms are released freely on certain platforms, therefore, other scientists can re-run the code to verify the results (Easterbrook 2014) and perform their own analysis upon them. Nowadays, there exist many easy-to-use source code platforms which not only allow hosting of software but also facilitate collaboration and version tracking (Goodman et al. 2014). Open code can make it easier for researchers to connect with one another by increasing the discoverability and visibility of one's work, which facilitates rapid access to software resources, and creates new opportunities to interact with and contribute to ongoing communal projects (McKiernan et al. 2016). Publishing the source code is also crucial to ensure research quality, reliability, and reproducibility (Bell 2017; Howison and Bullard 2016).

In spite of the importance and benefits, several studies have found that open code in the scientific community is problematic (Barnes 2010) though the practice of sharing code has been well promoted and encouraged (Shamir et al. 2013). On the other hand, the scientific community is facing a “reproducibility crisis” where more than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments (Baker and Penny 2016). The emerging problems call for a systematic investigation of code sharing behaviors of scientists and the role of open code in scholarly communication.

Many studies have investigated the impact of open science in scholarly communication. For example, open access articles were proved to receive significantly more citations than non-open access articles (Wang et al. 2015). The increase in the number of citation times for papers with data available was also demonstrated, by comparing the number of citations that papers with and without publicly shared data available (Leitner et al. 2016). Although all of these existing research assessed the effect of open science, still, efforts focus on the impact of open code are lacking. Specifically, the limitations of existing studies can be mainly reflected as follows: First, a longitudinal analysis of open code is missing. Second, the methods of sharing code and the accessibility of open code is less investigated and their effects on citation impact remain unknown. Third, citation impact is complex (Thelwall and Maflahi 2020), thereby a comprehensive understanding of the role that open code plays in broadening the citation impact is needed, such as the differences between early citation and long-term citation and knowledge flows among domains. Therefore, it is necessary to conduct detailed analytics to understand the role of open code as an increasingly common element in scholarly communication.

Although citations have been a controversial measure of research quality and performance, citations provide an objective and quantitative measure of credit and attention flows in science, because citations are directly related to the research output’s visibility and impact. Thus, we used citations as a proxy to examine how open code improve research visibility and impact. We also investigated open code sharing behaviors of researchers. In particular, this paper will focus on the field of computational linguistics with the following research questions:

- Q1: What is the trend of papers with open code in the field of computational linguistics?
- Q2: What are the effects of open code on citations?

This paper aims to answer these questions empirically through examining 12 years of fulltext papers from proceedings of Annual Meeting of the Association for Computational Linguistics (ACL) and Conference on Empirical Methods in Natural Language Processing (EMNLP). Both of them are premium publication venues of Computational Linguistics and Natural Language Processing (CL/NLP). Due to their representative roles within CL/NLP and similar reputations, we used papers from both for our study. Our dataset includes full text of 2,603 long papers published in the two conferences from 2006 to 2017.

## Method

### *Data collection*

The dataset used in this study contains open access, full-text papers from proceedings of Annual Meeting of the Association for Computational Linguistics (ACL) and Conference on Empirical Methods in Natural Language Processing (EMNLP). Both of them are premium publication venues of Computational Linguistic and Natural Language Processing (CL/NLP). We chose these two conferences because of their similar academic reputations and impacts, their nature of computation, and our domain knowledge. The access point to the dataset is provided by ACL Anthology and it is publicly available. We retrieved from ACL Anthology all papers in both conferences that were published between 2006 and 2017. Due to the nature of manual coding, we restricted the publication type as full paper. The final data collection includes 2,603 papers: 1,433 EMNLP conference papers, and 1,170 ACL conference papers, respectively. The metadata and citation data of papers were retrieved by Microsoft Academic. Microsoft Academic was proved to perform well in terms of both publication and citation coverage. The retrieved metadata of paper included the number of authors and number of references.

### *Coding procedures for identifying open code*

We manually identified the URLs that are for sharing code. We randomly selected a sample for developing a coding guideline. The basic rules of the guideline are: checking if a URL is one referring to a code repository and then checking if the code is contributed by the paper. Coders should read the URL and the context of the URL, visit the URL through a web browser, and read the code details when necessary.

We conducted a reliability test for our coding method. Two coders, both of whom are doctoral students in information science, coded independently on a sample of 200 randomly selected documents to annotate the URLs according to the guideline. To measure the inter-rater reliability (IRR) between the two coders, we calculated and achieved an IRR of 0.94, which was considered to provide sufficient reliability for one coder to code all the documents.

We found seven papers that each of them contains more than one shared code repositories. Due to the small amount of these special cases, we defined open code as a binary variable, i.e., having open code or not.

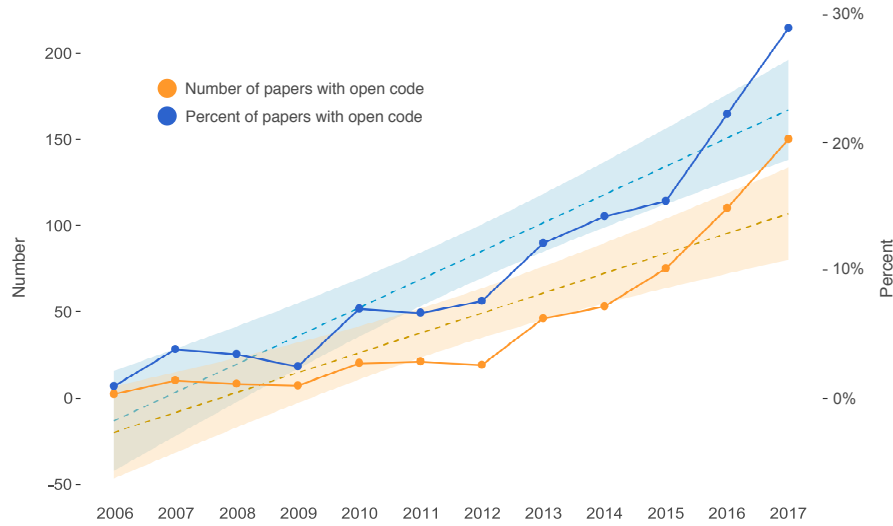
### *Examining accessibility*

The accessibility of an open code repository fundamentally determines its effectiveness in scholarly communication. It is necessary to distinguish the accessibility of open code repositories. To check the availability, we first checked the errors of extracted URLs to avoid wrongly marking URL as inaccessible ones due to conversion issues. We resolved the common errors in PDF-to-text conversion and data extraction, such as redundant whitespace, illegal characters in URLs, and missing of HTTP/HTTPS declaration. Once the link errors were corrected, the URLs were then automatically tested for their accessibility status using a Python package named `urllib`. Upon completion of the testing, we manually checked and tried to remedy the inaccessible URLs: First, the URLs were examined through browser visiting; Second, Google search was utilized as a tool in screening some unexpected errors that might be caused by missing characters in a URL for instance; Finally, the original full text (PDF) was used as a last reference for cross-checking.

## **Results**

### *Distribution of open code papers by year*

The distribution of open-code papers by publication year is shown in Figure 3. The distribution for the quantity of the open-code papers is visualized in orange and the percentage of the open-code papers for each year over all papers published in that year is in blue. The dashed lines represent the lines of best fit generated from linear regression, showing the temporal tendency of the distribution. The corresponding bands around the dashed regression lines indicate their confidence region from standard errors. The quantity of the open-code papers increased by nearly one third between 2006 and 2017. Specifically, after slow and fluctuating growth during 2006 to 2012, rising trends were observed in the quantity and percentage of the open-code papers from 2012 (19 open code papers, out of 7.75% of all papers) to 2017 (150 open code papers, out of 28.96% of all papers), especially large increase occurred in 2016 (110, 22.22%) and 2017 (150, 28.96%).



**Figure 1. Distribution of open code across years.**

### *Results of Negative Binomial regression models*

Many factors that have been proved to have a general influence on citation counts. To establish the adjusted covariation between open code and citation counts of open-code and non-open-code papers, we considered other factors by conducting a series of NB and ZINB models.

We analyzed citations of papers as outcome variable from four perspectives. At first, we examined effects of open code on the total number of citations up to the end of April in 2019. Second, we used three fixed citation windows of one, three, and six years. Citations within a fixed window is denoted as  $\text{citations}(w)$  where  $w$  is the size of a citation window.  $\text{citations}(w)$  of a paper  $p$  is the number of papers cited  $p$  and published within  $[y_p, y_p + w]$  (including  $y_p$  and  $y_p + w$ ) where  $y_p$  is the publication year of  $p$ . Using multiple citation windows can not only examine short- and long-term citations, but also test the reliability of the results of citation analysis (Harnad 2009). Third, we examined how the effects of open code changed over time. The popularity of sharing open code increased dramatically in recent years, which may lead to varying effects of open code.

Previous scientometric studies have demonstrated various factors that have influences on citation counts. We measured conference-level, paper-level, and author-level characteristics that may have influences on citation counts. An overview of definitions for the main independent variables is provided in Table 1. It worth noting that highly cited authors for each year were identified separately. For each year, we collected 10-year citations of authors before the year and only authors who published in the year were considered. For example, for identifying top 1% highly cited authors in 2001, we first selected the authors published papers in these two conferences in 2001 and then collected citations of these authors from 1991 to 2000 to identify top 1% cited authors.

We built Negative Binomial (NB) and Zero-Inflated Negative Binomial regression models to validate the effects of open code on citations. Both models were commonly used to model citation count. A ZINB model include a NB model for citation count and a logit model for modeling excess zero citations. We conducted a Vuong test to learn which model is superior for each dataset and chose the superior one to conduct our analysis. NB models using log as the link function are defined as follow:

$$\text{citations}(w) \sim \text{code} + \text{year} + \text{conference} + \text{reference} + \text{author} + \text{top1} + \text{top5} + \text{top10}$$

**Table 1 Does sharing open code get more future citations? Each column is the coefficients for open code and other control variables from a Negative Binomial Regression model of open code on future citation counts; p values are reported in parentheses.**

	<i>Model 1</i>	<i>Model 2</i>	<i>Model 3</i>	<i>Model 4</i>
<i>code</i>	0.653 (0.000)	0.665 (0.000)	0.700 (0.000)	0.278 (0.000)
<i>year</i>			0.209 (0.000)	0.209 (0.000)
<i>conference</i>			0.018 (0.675)	0.013 (0.751)
<i>reference</i>				
<i>author</i>				
<i>top1%_author</i>				
<i>top5%_author</i>				
<i>top10%_author</i>				
Dispersion parameter	0.552	0.542	0.509	0.462
2*log-likelihood	-38,656	-38,754	-39,096	-39,618
AIC	38,677	38,768	39,107	39,624
Vuong test (NB >ZINB)	33.048 (0.000)	34.969 (0.000)	33.925 (0.000)	37.464 (0.000)
N	4,098	4,098	4,098	4,098

Table 1 reports results from NB regressions of future citation counts on open code, with different controls. Model 1 reports, without any control variables, the percentage change in the number of citations and papers with open code. We found that sharing open code is statistically related to future citations; our estimated coefficients indicate that sharing open code is associated with 27.8% more citations ( $p < 0.001$ ). There are potential concerns with interpreting the unadjusted relationship between open code and citations as a measure of open code's value. Older papers published in earlier years have more time to accrue citations. Besides the publication years, ACL usually takes place earlier than EMNLP, thus ACL papers tend to have more time for being cited.

Model 2 addresses these concerns by including cohort effects by publication years and conferences. The inclusion of these cohort effects means that our estimates are based only on comparisons of having open code or not and citations of papers published in both the same year and the same conference. Controlling for cohort effects reinforce our finding based on results of Model 1. We expect a 70.0% increase in citations ( $p < 0.001$ ) by sharing open code.

However, some studies have demonstrated that bibliometric factors of a paper can have influence on its citation counts. Papers with more references and authors tends to be more frequently cited. Model 3 adds controls describing a paper's bibliometric characteristics, including the number of authors and the number of references, to examine if open code factor can have stronger predictive effects than these well studied bibliometric factors. Controlling for paper characteristics slightly attenuates the relationship between open code and citations but the relationship is still strong: having open code in a paper is associated with a 65.5% increase in future citations.

Citations could also be partly explained by Matthew effect in science and "success-breeds-success" phenomenon. Papers by established authors whose papers have been highly cited in

the past can be expected to be more frequently cited than papers by less established authors, regardless of the true quality of their work.

Model 4 controls for the citation history of a paper's authors. Specifically, we used three variables to describing the number of top 1%, 5%, and 10% authors that are highly cited in the past ten years. Including these variables, our estimates remain stable: papers with open code tend to have 65.3% more citations than papers without open code. The statistically significant effects of other factors are inconsistent with existing bibliometric studies of factors affecting citation counts: (a) Papers with more references and authors tend to have more citations. (b) Papers by more highly cited authors are expected to have more citations. Comparing the estimates in Model 4, open code has been shown stronger effects on future citations of a paper than other control factors. For example, having open code exerts a positive effect on citations of a paper nearly equivalent to that of co-authorship with two top-1% highly cited authors.

## Conclusion and future work

By measuring the effects of open code on citation impact over time, we found that the papers with open code have more citations than papers without open code on average for most of years. We have also applied two regression models on the investigation of the effects of open code over various datasets. The ZINB regression model with the variables is found to result in inferior fit with the data, while statistically significant relationship is found between open code and citation impact when using the NB regression model.

We will compare the effects of open code between long-term citations and short-term citations, and between intradomain citations and inter-domain citations. We will also examine the accessibility of the open code and if accessibility would lead to different effects on citation.

## References

- Baker, M., & Penny, D. (2016, May 26). Is there a reproducibility crisis? *Nature*. Nature Publishing Group. <https://doi.org/10.1038/533452A>
- Barnes, N. (2010, October 14). Publish your computer code: It is good enough. *Nature*. <https://doi.org/10.1038/467753a>
- Bell, V. (2017). Open science in mental health research. *The Lancet Psychiatry*, 4(7), 525–526. [https://doi.org/10.1016/S2215-0366\(17\)30244-4](https://doi.org/10.1016/S2215-0366(17)30244-4)
- Easterbrook, S. M. (2014). Open code for open science? *Nature Geoscience*. <https://doi.org/10.1038/ngeo2283>
- Goodman, A., Pepe, A., Blocker, A. W., Borgman, C. L., Cranmer, K., Crosas, M., et al. (2014). Ten Simple Rules for the Care and Feeding of Scientific Data. *PLoS Computational Biology*, 10(4), e1003542. <https://doi.org/10.1371/journal.pcbi.1003542>
- Harnad, S. (2009). Open access scientometrics and the UK Research Assessment Exercise. *Scientometrics*, 79(1), 147–156. <https://doi.org/10.1007/s11192-009-0409-z>
- Haustein, S. (2016). Grand challenges in altmetrics: heterogeneity, data quality and dependencies. *Scientometrics*, 108(1), 413–423. <https://doi.org/10.1007/s11192-016-1910-9>
- Howison, J., & Bullard, J. (2016). Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. *Journal of the Association for Information Science and Technology*, 67(9), 2137–2155. <https://doi.org/10.1002/asi.23538>
- Leitner, F., Bielza, C., Hill, S. L., & Larrañaga, P. (2016). Data publications correlate with citation impact. *Frontiers in Neuroscience*, 10, 419. <https://doi.org/10.3389/fnins.2016.00419>
- McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., et al. (2016). How open science helps researchers succeed. *eLife*, 5(JULY), 1–19. <https://doi.org/10.7554/eLife.16800>
- Piwovar, H. (2013). Value all research products. *Nature*, 493(7431), 159–159. <https://doi.org/10.1038/493159a>
- Shamir, L., Wallin, J. F., Allen, A., Berriman, B., Teuben, P., Nemiroff, R. J., et al. (2013). Practices in source code sharing in astrophysics. *Astronomy and Computing*, 1, 54–58.

<https://doi.org/10.1016/j.ascom.2013.04.001>

- Thelwall, M., & Maflahi, N. (2020). Academic collaboration rates and citation associations vary substantially between countries and fields. *Journal of the Association for Information Science and Technology*, 71(8), 968–978. <https://doi.org/10.1002/asi.24315>
- Wang, X., Liu, C., Mao, W., & Fang, Z. (2015). The open access advantage considering citation, article usage and social media attention. *Scientometrics*, 103(2), 555–564. <https://doi.org/10.1007/s11192-015-1547-0>