How data publishing work is credited in group research?

AUTHORS SECTION

Jiang, Tianji (of 1st Author) University of California, Los Angeles, United States | tianji008@ucla.edu

Chen, Yuejiao (of 2nd Author) University of Illinois Urbana-Champaign, United States | cjiao4@illinois.edu

ABSTRACT

This study provides quantitative evidences on researchers' lack of incentives on data publishing work. Our findings will facilitate the efforts on data opening, and arouse more attention to the data publishing work. More importantly, our study will also help researchers and policy makers to design metrics for measuring and rewarding the contribution that data sharing makes, which will benefit the academia forever.

KEYWORDS

Scientific data; Data sharing; Data paper

SECTIONS

"Data sharing" generally refers to the act of releasing data in a form that can be used by other individuals (Pasquetto et al., 2017). Promoting data sharing has been the consensus of our academic community today, as it allows people to benefit from current researches in a new approach. Data sharing has been credited with increasing efficiencies in research, more reproducible science, maximizing the use of a valuable resource (Walport & Brest, 2011), allowing for an expansion of innovation, escalating collaboration (Popkin, 2019), and has been credited with the rapid development of COVID-19 vaccines, therapies and diagnostics (Staunton et al., 2021). It is indicated that data sharing encompasses opening the dataset itself but also releasing its key context information, including how it is collected, processed, preserved and shared. There are various means of sharing data, such as posting datasets on researchers' or laboratory websites; depositing datasets in publicly accessible collections; and attaching data as supplemental materials in journal articles (Wallis et al. 2013). Yet these practices are criticized for deficiency in sharing the contextual information of dataset, which is necessary for researchers to repurpose the dataset collected by the others (Culina et al., 2018). A relatively remedy for the problem in many fields is to disseminate a dataset as a "data paper." Data paper is an article that provides descriptions of methods for collecting, processing, and verifying data (Ahn et al. 2012), which improves data provenance and reusability.

Publishing data gives researchers extra credits. Previous studies have found that papers with publicly available datasets receive a higher number of citations than similar studies without available data (Piwowar & Vision, 2013). Nevertheless, making research data publicly available also has challenges and costs. Enabling data reuse through publishing and describing the data is a costly proposition in and

of itself, and some scholars believe that these works often fail to receive deserved recognitions in scholarship (Popkin, 2019).

It is a massive undertaking to understand why researchers lack incentives to share data and address it. In this study, we take a small step towards big question by exploring how data publishing work is credited by researchers themselves in a group study. We retrieved 3705 data papers from Data in Brief, a multidisciplinary and peer-reviewed journal which publishes data papers and provides access to research data. We also retrieved 3705 research articles where the research data are created. Each data paper refers to a research article. Considering that the contributors' names listed as authors and the sequence of authors usually reflect contributions to the work accessed by the research group (Traditionally, the first author contributes most, whereas the position of subsequent authors is usually decided by contribution), we compare the authors of each data paper and the relevant research paper to answer our question.

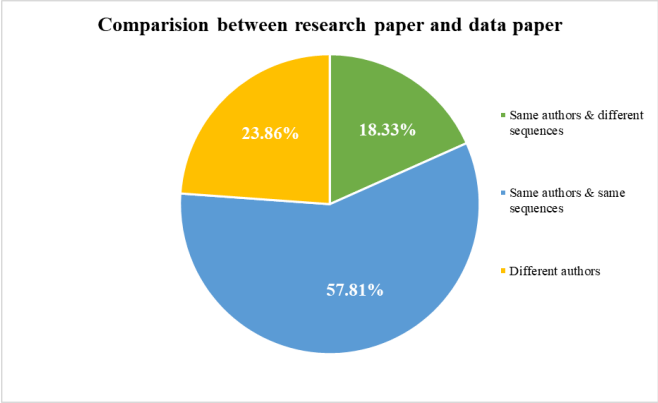

Figure- 1

Among these papers, 2142 pairs of data paper and research paper are just the same in author information, whereas 1563 pairs are different (As shown in Figure-1). Moreover, 679 pairs of paper have the same authors in different sequence while the rest have different authors.

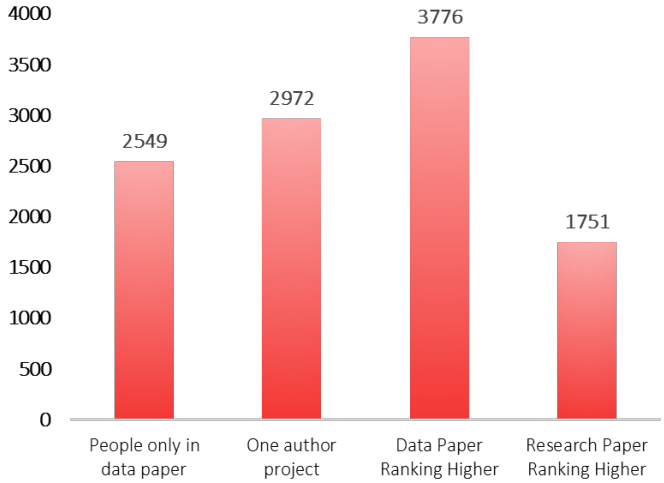

Figure- 2

Among the 1563 pairs papers that are different in author information, we noticed that research paper has more or the same number of authors than its relevant data papers in most cases (in 1353 pairs). The means of authors' number of data paper is 5.86, while the relevant research paper is 7.59. It confirms

that not all research participants will contribute to data work in group research, which lays the foundation of our following analysis. Furthermore, we count the data paper authors' rankings in both research papers and data papers. Excluding the 2972 authors who are the only author of the research paper and data paper, and the 2549 authors who are only listed as author of the data paper, 3776 authors get a higher ranking in data paper than in research paper, while only 1751 get a higher ranking in research paper than in data paper (As shown in Figure-2).

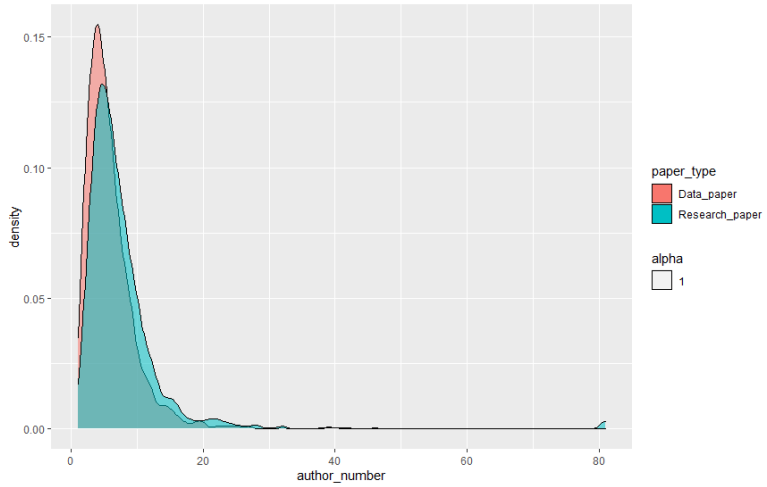

Figure- 3

Besides, among the papers that are different in author information, the first author of data paper is also listed as the top 3 authors in the relevant research paper in 1266 pairs (972 as first author), which indicates that the research participants who contribute much to data publishing are often given significant credits by the research group. However, we cannot conclude that data publishing work is credited a lot by researchers today. The main authors of a research paper are often the ones who design the study and contribute the most to data collection and analysis, and thus they are often still listed as top authors of a data paper even if they make little contributions to the data publishing work. Further studies are necessary to clarify it.

We also try to quantify the difference between author information the pairs of data paper and research paper from two perspective, the authors and the sequence. We calculate the Jaccard similarity coefficient to measures the similarity between the author information of the pairs of papers to see how many authors are shared and distinct. The average Jaccard index is 0.18, which indicates that data papers and their relevant research papers are quite similar in authors. We then use LCS (longest common subsequence) to measure the similarity in sequence. The average LCS of the pairs of papers that are different in author information is 0.33, which indicates that the author positions are usually different between data papers and their relevant research papers. In addition, we notice that the authors that rank low in research paper appear as authors of the relevant data paper in many cases, which implies that the research group usually give little credits to the data publishing work, and thus only the marginal participants of the research would like to spend time on it.

$$d_J(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}.$$

Figure- 4

Finally, we quantify the difference between author information by counting the total number of each case where an author gets a specific ranking in data paper and a specific ranking in research paper. For instance, the author is listed as the second author in 194 research papers, the third author in 62 papers, and the fourth author in 45 papers when he or she ranked as the first author of a data paper. We visualize the result as Figure-5. Furthermore, we calculate the average ranking difference of a data paper's 1st to 9th author (as shown in Figure-6). According to the results, an data paper author tends to get a lower ranking in research paper when he or she gets a high ranking (top 6) in data paper, while he or she tends to get a higher ranking in research paper when receiving a low ranking in data paper. Considering that the top authors of a data paper usually are the main contributors of the data publishing work, the results may provide evidences for the arguments that data publishing work doesn't receive significant credits in today's practices.

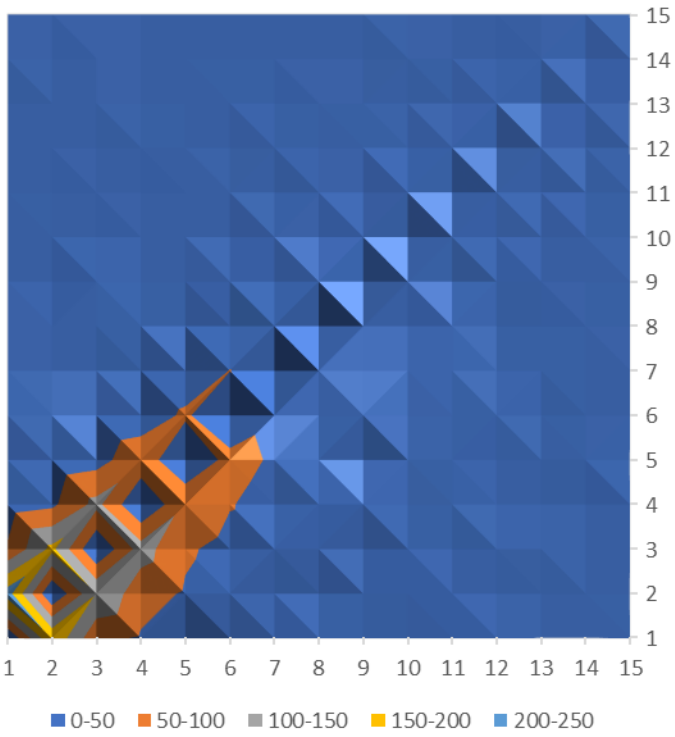

Figure- 5

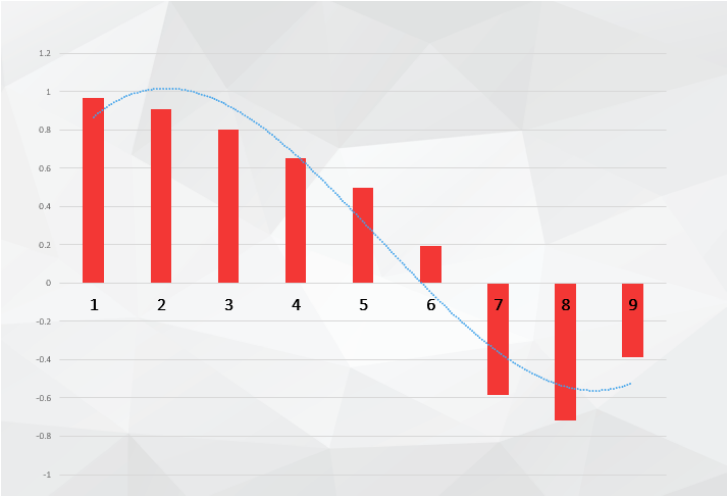

Figure- 6

In the next step, we will follow up by studying the difference in detail, and try to draw a whole picture of how data publishing work is credited. This study provides quantitative evidences on researchers' lack of incentives on data publishing work. Our findings will facilitate the efforts on data opening, and arouse more attention to the data publishing work. More importantly, our study will also help researchers and policy makers to design metrics for measuring and rewarding the contribution that data sharing makes, which will benefit the academia forever.

REFERENCE

Ahn, C P Alexandroff, R Allende Prieto, C Anderson, S F Anderton, T Andrews, B H et al. (2012). The Ninth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the SDSS-III Baryon Oscillation Spectroscopic Survey. Astrophysical Journal 203: 21. https://doi.org/10.1088/0067-0049/203/2/21

Culina, A., Crowther, T. W., Ramakers, J. J. C., Gienapp, P., & Visser, M. E. (2018). How to do meta-analysis of open datasets. Nature Ecology & Evolution, 2(7), 1053–1056. https://doi.org/10.1038/s41559-018-0579-2

Pasquetto, I., Randles, B., & Borgman, C. (2017). On the Reuse of Scientific Data. Data Science Journal, 16(0), 8. https://doi.org/10.5334/dsj-2017-008

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. PeerJ, 1, e175. https://doi.org/10.7717/peerj.175

Popkin, G. (2019). Data sharing and how it can benefit your scientific career. Nature, 569(7756), 445–447. https://doi.org/10.1038/d41586-019-01506-x

Staunton, C., Barragán, C. A., Canali, S., Ho, C., Leonelli, S., Mayernik, M., Prainsack, B., & Wonkham, A. (2021). Open science, data sharing and solidarity: who benefits? History and Philosophy of the Life Sciences, 43(4), 115. https://doi.org/10.1007/s40656-021-00468-6

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. PLoS ONE, 8(7), e67332. https://doi.org/10.1371/journal.pone.0067332

Walport, M., & Brest, P. (2011). Sharing research data to improve public health. The Lancet, 377(9765), 537–539. https://doi.org/10.1016/S0140-6736(10)62234-9