The Science of Scientific Software

Eva Maxfield Brown

Background

The Science of Science (SciSci) is an interdisciplinary field dedicated to studying the organization of science and its economic, epistemic, and societal impact. Research in SciSci therefore seeks to understand the practices - both social and technical - that lead to new knowledge production (1). For example, research in the Science of Science has focused on understanding demographic shifts and changes of scientists (i.e. who is working on what topics), how basic and applied research is funded, the economic impact of science, and famously how collaboration occurs– between whom, and to what effect (2–7). The primary method of analysis for many Science of Science investigations has so far been quantitatively analyzing large corpora of well structured, readily available, scientific publications. However, analyzing only the final product of scientific endeavor overlooks a crucial element of contemporary science: its computational nature (8–10).

To better understand how science is practiced today, SciSci researchers must look at more than just the final product of research. We need to delve into the notes, documentation, and importantly, the computational scripts that drive research. In short, software permeates virtually all scientific disciplines (8–12). By studying scientific software, we gain a deeper understanding of the collaborative processes, technical dependencies, and digital supply chains that underpin contemporary scientific knowledge production.

Proposed Research

This proposal aims to leverage traditional Science of Science methods to explore the specific domain of scientific software. By combining analysis of publications with detailed information extracted from software repositories and source code, this research will address questions such as:

- How are scientific software tools built and maintained?
- Who are the key players in the scientific software ecosystem?
- Do critical pieces of software exist and how can we ensure their sustainability?

Drawing upon existing qualitative and quantitative research, my work seeks to move beyond analyzing solely the textual mentions of software within publications, and instead directly analyze research source code and repository metadata alongside linked research publications (9, 10, 13–17). In doing so, we can revisit previous topics in SciSci anew:

- Identification: Ask who is involved in knowledge production, the roles they play, and how these roles have changed over time.
- Collaboration: Look beyond co-authorship and acknowledgements to understand the direct and indirect dependencies of scientific production.
- Economic: Understand how investments in basic and applied research translate to new general purpose computational tools which have financial and societal impacts.

My proposed research will leverage machine learning, natural language processing (NLP), and network analysis methods. I have already developed and evaluated custom machine learning models for scientific software identification, and I am actively applying network analysis methods to understand the landscape of research software development (18–21). Moving forward, I plan on creating efficient and accurate methods for the bi-directional linkage of publications and their associated code repositories. With these links we can then employ network analysis to compare both authorship and dev-contributions to scientific research.

The proposed research topics and methods also draw upon my own experience in creating and maintaining scientific software packages (AICSImageIO, Speakerbox, Council Data Project), and both building and sustaining open-source communities (22, 23). I believe this provides invaluable first-hand insights into the complexities of scientific software development and collaboration.

Contribution

My research directly addresses the need for a deeper understanding of scientific software development and its impact, ultimately contributing to the following communities:

- Science of Science Scholars
- Information Scientists
- Technology Policy Experts

This work will create new sources of data, models, and methods, which help illuminate research trends, collaboration patterns, and computational knowledge diffusion and utilization. More broadly, if successful this work will positively impact the scientific enterprise– allowing public funding agencies to better support the long-term sustainability of computationally dependent research and allow others to gain deeper insight into scientific process and practice.

Bibliography

1. Santo Fortunato, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Stasa Milojevic, Alexander Michael Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, and Albert-László Barabási. Science of science. *Science*, 359, 2018.
2. Mark E. J. Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America*, 101:5200 – 5205, 2004.
3. Stasa Milojevic. Quantifying the cognitive extent of science. *J. Informetrics*, 9:962–973, 2015.
4. Carole J. Lee, Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. Bias in peer review. *J. Assoc. Inf. Sci. Technol.*, 64:2–17, 2013.
5. Jevin D. West, Jennifer Jacquet, Molly M. King, Shelley J. Correll, and Carl T. Bergstrom. The role of gender in scholarly authorship. *PLoS ONE*, 8, 2012.
6. Noriko Hara, Paul Solomon, Seung-Lye Kim, and Diane H. Sonnenwald. An emerging view of scientific collaboration: Scientists' perspectives on collaboration and factors that impact collaboration. *J. Assoc. Inf. Sci. Technol.*, 54:952–965, 2003.
7. Adèle Paul-Hus, P. Mongeon, Maxime B. Sainte-Marie, and Vincent Larivière. Who are the acknowledgees? an analysis of gender and academic status. *Quantitative Science Studies*, 1:582–598, 2020.
8. Demitri Muna, Michael Alexander, Alice Allen ..., and Andrea Zonca. The astropy problem. *arXiv: Instrumentation and Methods for Astrophysics*, 2016.
9. Prakash Prabhu, Thomas B. Jablin, Arun Raman, Yun Zhang, Jialu Huang, Hanjun Kim, Nick P. Johnson, Feng Liu, Soumyadeep Ghosh, Stephen R. Beard, Taewook Oh, Matthew Zoufaly, David Walker, and David I. August. A survey of the practice of computational science. *2011 International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, pages 1–12, 2011.
10. Jo Erskine Hannay, Hans Petter Langtangen, Carolyn MacLeod, Dietmar Pfahl, Janice Singer, and Greg Wilson. How do scientists develop and use scientific software? *2009 ICSE Workshop on Software Engineering for Computational Science and Engineering*, pages 1–8, 2009.
11. Joel Cohen, Daniel S. Katz, Michelle Barker, Neil Philippe Chue Hong, Robert Haines, and Caroline Jay. The four pillars of research software engineering. *IEEE Software*, 38:97–105, 2020.
12. Pauli Virtanen, Ralf Gommers, Travis E. Oliphant ..., and Y. Vázquez-Baeza. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, 17:261 – 272, 2019.
13. Jeffrey C. Carver, Nic Weber, Karthik Ram, Sandra Gesing, and Daniel S. Katz. A survey of the state of the practice for research software in the united states. *PeerJ Computer Science*, 8, 2022.
14. Cai Fan Du, Johanna Cohoon, Patrice Lopez, and James Howison. Softcite dataset: A dataset of software mentions in biomedical and economic research publications. *Journal of the Association for Information Science and Technology*, 72:870 – 884, 2021.
15. Juan Pablo Alperin, Lesley A. Schimanski, Michelle La, Meredith T. Niles, and Erin C. McKiernan. The value of data and other non-traditional scholarly outputs in academic review, promotion, and tenure in canada and the united states. *The Open Handbook of Linguistic Data Management*, 2022.
16. Terry Bucknell. Recognising influence: helping authors of non-traditional research outputs evidence the reach and potential impacts of their work. 2016.
17. Ana-Maria Istrate, Donghui Li, Dario Taraborelli, Michaela Torkar, Boris Veytsman, and Ivana Williams. A large dataset of software mentions in the biomedical literature. *ArXiv*, abs/2209.00693, 2022.
18. Eva Maxfield Brown, Lindsey Schwartz, Richard Lewei Huang, and Nicholas M. Weber. Soft-search: Two datasets to study the identification and production of research software. *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 228–231, 2023.
19. Andrew Nesbitt. Package and Dependency Metadata for CZI Hackathon: Mapping the Impact of Research Software in Science, October 2023.
20. Eva Maxfield Brown. A Dependency Graph for 460,000 Papers and Their Software Mentions from the CZI Software Mentions Dataset, October 2023.
21. Eva Maxfield Brown. Research software graph, 2023.
22. Eva Maxfield Brown, To Huynh, Isaac Na, Brian Ledbetter, Hawk Ticehurst, Sarah Liu, Emily Gilles, Katlyn Greene, Sung Cho, Shak Ragoler, and Nicholas Weber. Council data project: Software for municipal data collection, analysis, and publication. *J. Open Source Softw.*, 6:3904, 2021.
23. Eva Maxfield Brown, To Huynh, and Nicholas Weber. Speakerbox: Few-shot learning for speaker identification with transformers. *J. Open Source Softw.*, 8:5132, 2023.