Complex Systems & Data Science

Honors Thesis Proposal

Text

Description automatically generated

Uncovering Scientific GitHub: Analyzing Links Between Academic Papers and GitHub Repositories

Alexander P. Friedrichsen, James P. Bagrow, Ph.D, and Laurent Hébert-Dufresne, Ph.D.

October 29, 2021

Contents

[Background 2](#_Toc86402902)

[Motivation 3](#_Toc86402903)

[Related Literature 3](#_Toc86402904)

[Obtaining a sample: 3](#_Toc86402905)

[Studies and analysis on traceability of scientific artifacts: 4](#_Toc86402906)

[Specific Work Plan 5](#_Toc86402907)

[References 7](#_Toc86402908)

# Background

Open science and the use of open source software development platforms like GitHub fit a model of innovation that combines private and collective action. This new paradigm of social coding and free access to information provides society the best of both organizational worlds [6]. Open science has begun to receive a spotlight in research, as researchers question the true impact of free information and the ins and outs of current platforms and practices surrounding it. GitHub is one such platform; GitHub is the predominant collaborative coding site with over 100 million repositories as of November 2018 [14]. It is also the primary data source for researchers seeking to understand how users collaborate on software projects. [7]. Several features on GitHub serve to promote teamwork and discussion within a project and allow users to track and clone (fork) others’ work. GitHub is an obvious and compelling choice for a data source in any study pertaining to open source science and software development.

Reproducibility is an ever-present concern in science, and with the ever-increasing flow of scientific papers being published, reproducibility and new studies that build upon prior research are paramount. Reproducibility is the critical benchmark of published science by which results are validated or refuted [12]. Furthermore, there is a reproducibility crisis that may be assuaged by access to researchers’ raw data and code [9]. One way to increase reproducibility is linking between papers and the code behind their conclusions. Teams in numerous academic fields sometimes publish code used in their research papers on GitHub in addition to publishing the papers themselves and often include links between the paper and repository or vice versa. This practice coincides with the growth of open science yet is not standardized and is not yet fully understood. Measuring the quantity and clarity of links between artifacts is the first step to understanding and improving the traceability of scientific artifacts.

# Motivation

We seek to identify a dataset of (bidirectionally) linked repositories and papers to uncover the potential indirect effects of papers on repositories. We will examine the distribution of original repository citations from papers that cite a linked paper. We intend to analyze how papers cite repositories and other papers, and whether this relationship changes for different fields of study and popularity of the cited artifact. Finally, we intend to examine the evolution of how papers have began citing repositories in addition to other papers. During our analysis we hope to find additional questions that explain these sought indirect effects.

# Related Literature

## Obtaining a sample:

Using a sample from GitHub has many pros: the convenience of the GitHub Archive as a data source, the large quantity and variety of information, and importantly prior work that helps avoid pitfalls in working without GitHub data. Kalliamvakou *et al.* [7] help researchers identify some of these pitfalls, such as knowing not all repositories are full projects or exist on GitHub in combination with some other platform. It is important to fully understand the sample one chooses to work with in all its facets. More specific to our stated research question is the selection of a sample of papers traceable to repositories. In two related studies, to obtain a sample of 20000 README files, the authors used regular expressions to match paper-link-appearing text objects [5, 17]. Then, a representative sample was taken from within this 20000 and links were tested for validity. This data set is available for future use [3]. Entitled “paper2repo” a GitHub repository recommender uses machine learning to recommend users with repositories that are like a chosen paper on an academic search system such as Microsoft Academic [13]. However, this doesn’t necessarily give the repositories used by the authors of the paper itself, which is problematic for the type of analysis we want to carry out. Another method to obtain a sample was to take all links from papers to repositories from Microsoft Academic Graph classified as primary and belonging to computer science, resulting in collection of around 5000 at time of publication links, of which about three-fifths were downloadable [2]. An effort aimed at collecting machine learning projects and their related papers has been cultivated on the site [paperswithcode.com](https://paperswithcode.com/) [10]. This site is another potential source for accurate links between repositories and papers.

## Studies and analysis on traceability of scientific artifacts:

Our research question falls squarely in the emerging zone of traceability of scientific artifacts. Much of the prior research here also looks at links inside of GitHub. Prana *et al.* [11] categorize the contents of README files systematically by first manually labelling and then automatically classifying different sections of README files. One of the sections labelled “References” includes external links out of the repository which is of interest when trying to identify links to papers from repositories. The aptly titled paper by Hata *et al.* is one of the first to explore links between papers and repositories. The software domain was categorized, evolution of links analyzed (link maintenance), and potential bidirectional links categorized. These bidirectional links are of particular interest, as they are closely related to one of our potential metrics for indirect influence. This bidirectionality was categorized by including links to papers that do and do not link back to their own repository, broken links (404 errors), and papers that do not link back to any repository. Links from academic papers link to GitHub most of the time, among all types of links [17].

Numerous other relevant questions were investigated in Wattanakrienkrai *et al.* [17]. The number of repositories that link to papers has been estimated [17]. The authors of linked papers and repositories have been examined by affiliation and found to come mostly from academia (Milewicz, 17]. In addition, the authors have been cross compared with repository contributors to see if the people publishing the paper are the same making edits to the code and to see if the lead authors are the lead contributors to the repository, finding that 40% of repositories are being edited by the authors of the linked paper, and greater than 50% are other contributors implementing the research published in the paper. The most referenced arXiv papers from GitHub repositories are identified using a bipartite network – the most cited papers are also the most referenced in GitHub. Finally, the evolution of links on GitHub is also analyzed finding that changes in links are rare. Additionally, when a paper is updated a corresponding update to the link in a linked repository is rare.

Several questions about the distribution of basic properties of linked repositories were examined by Färber [2]. The distribution of repositories by number of stars, forks, contributors, lengths of repository manuals, programming languages used, machine learning frameworks used, and fields of study were each visualized. Moreover, affiliation by institution, conference series, and journal was also displayed. Finally, it is important to note the existence of multiple studies that have attempted to identify the impact of data sharing and public availability of data on article popularity and citations. Publicly available code is found to increase citation counts on papers in the machine learning field [16]. Publicly available articles in the biostatistics field with open access to code were found to have a 2-fold increase in number of citations [15]. Higher biostatistics article citations were associated with higher number of references [15]. Data sharing policies implemented by journals did not appear to lead to higher citation counts in the following five years after enacting the policies [Christensen].

# Specific Work Plan

* October 1st – November 1st
  + Expand literature review
  + Submit proposal and receive confirmation
* November 1st – January 17th
  + Confirm defense committee members
  + Finish data engineering
  + Preliminary visualizations
  + Meetings over winter break to discuss progress (?)
  + Sign up for 3 more credits of thesis work in the spring
* January 17th – February 4th
  + Finished most coding necessary
  + Finish visualizations
  + Begin writing
* February 11th
  + Finish writing first (personal) draft of thesis, without figures/bibliography
* March 12th
  + Complete first draft due
* April 1st
  + Final draft due
  + Announce defense and date to public (April 11th?)
* April 11-15th
  + Hold defense
  + Final edits and feedback revisions if necessary
* April 30th
  + Manuscript due
* May
  + Publish?

# References

[1] Christensen, G., Dafoe, A., Miguel, E., Moore, D. A., & Rose, A. K. (2019). A study of the impact of data sharing on article citations using journal policies as a natural experiment. *PLOS ONE*, *14*(12), e0225883. <https://doi.org/10.1371/journal.pone.0225883>

[2] Färber, M. (2020). Analyzing the GitHub Repositories of Research Papers. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, 491–492. <https://doi.org/10.1145/3383583.3398578>

[3] *GitHub Repositories with Links to Academic Papers: Open Access, Traceability, and Evolution*. (2020). Software Engineering Lab @ NAIST. <https://github.com/NAIST-SE/GH2Papers> (Original work published 2020)

[4] *Google Scholar*. (n.d.). Retrieved October 27, 2021, from <https://scholar.google.com/?inst=11539667347833751441>

[5] Hata, H., Guo, J. L. C., Kula, R. G., & Treude, C. (2021). Science-Software Linkage: The Challenges of Traceability between Scientific Knowledge and Software Artifacts. *ArXiv:2104.05891 [Cs]*. <http://arxiv.org/abs/2104.05891>

[6] Hippel, E. von, & Krogh, G. von. (2003). Open Source Software and the “Private-Collective” Innovation Model: Issues for Organization Science. *Organization Science*, *14*(2), 209–223. <https://doi.org/10.1287/orsc.14.2.209.14992>

[7] Kalliamvakou, E., Gousios, G., Blincoe, K., Singer, L., German, D. M., & Damian, D. (2016). An in-depth study of the promises and perils of mining GitHub. *Empirical Software Engineering*, *21*(5), 2035–2071. <https://doi.org/10.1007/s10664-015-9393-5>

[8] Milewicz, R., Pinto, G., & Rodeghero, P. (2019). Characterizing the Roles of Contributors in Open-Source Scientific Software Projects. *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*, 421–432. <https://doi.org/10.1109/MSR.2019.00069>

[9] Miyakawa, T. (2020). No raw data, no science: Another possible source of the reproducibility crisis. *Molecular Brain*, *13*(1), 24. <https://doi.org/10.1186/s13041-020-0552-2>

[10] *Papers with Code—The latest in Machine Learning*. (n.d.). Retrieved October 28, 2021, from <https://paperswithcode.com/>

[11] Prana, G. A. A., Treude, C., Thung, F., Atapattu, T., & Lo, D. (2019). Categorizing the Content of GitHub README Files. *Empirical Software Engineering*, *24*(3), 1296–1327. <https://doi.org/10.1007/s10664-018-9660-3>

[12] Ram, K. (2013). Git can facilitate greater reproducibility and increased transparency in science. *Source Code for Biology and Medicine*, *8*(1), 7. <https://doi.org/10.1186/1751-0473-8-7>

[13] Shao, H., Sun, D., Wu, J., Zhang, Z., Zhang, A., Yao, S., Liu, S., Wang, T., Zhang, C., & Abdelzaher, T. (2020). paper2repo: GitHub Repository Recommendation for Academic Papers. *Proceedings of The Web Conference 2020*, 629–639. <https://doi.org/10.1145/3366423.3380145>

[14] Thank you for 100 million repositories. (2018, November 8). *The GitHub Blog*. <https://github.blog/2018-11-08-100m-repos/>

[15] Wahlquist, A. E., Muhammad, L. N., Herbert, T. L., Ramakrishnan, V., & Nietert, P. J. (2018). Dissemination of novel biostatistics methods: Impact of programming code availability and other characteristics on article citations. *PLOS ONE*, *13*(8), e0201590. <https://doi.org/10.1371/journal.pone.0201590>

[16] WANG Shuwen, X. Y., & WANG Shuwen, X. Y. (2021). Influence Mechanism of Code-Sharing on Paper Citations：An Empirical Analysis on Computer Science Field. *Frontiers of Data and Computing*, *3*(2), 93–102. <https://doi.org/10.11871/jfdc.issn.2096-742X.2021.02.011>

[17] Wattanakriengkrai, S., Chinthanet, B., Hata, H., Kula, R. G., Treude, C., Guo, J., & Matsumoto, K. (2021). GitHub Repositories with Links to Academic Papers: Public Access, Traceability, and Evolution. *ArXiv:2004.00199 [Cs]*. <http://arxiv.org/abs/2004.00199>