Git/GitHub, Transparency, and Legitimacy in Quantitative Research

Zachary M. Jones

Published in *The Political Methodologist*, Fall 2013

The decreasing cost of computing power and the increase in the availability of a variety of public data has made quantitative research much more attractive to quantitative social scientists. I would argue that the increasing availability of public data and the availability of enormous computational power has generally been a good thing for the discipline. It has not been without cost however. Since researchers now have enormous flexibility in data collection and manipulation, as well as model selection, estimation, and reporting, it is often difficult to evaluate the internal and external validity of published findings. In other disciplines (notably psychology and medicine) there has been a perceived and actual increase in the false-positive rate of published quantitative research (Simmons, Nelson and Simonsohn 2011). Increases in the actual and/or percieved false-positive rate may have policy implications, as politicians and grant-giving institutions decide how to allocate limited resources.

What is the most effective way to deal with these issues? Many of the most prominent journals in political science now require replication materials. Replication archives are still not ubiquitous, and progress in this area is an important step towards decreasing the actual and perceived false-positive rate. There is also evidence of a disconnect between requirements and practice, at least in economics (Andreoli-Versbach and Mueller-Langer 2013). There has been a broader push to make research more transparent. For example, a recent issue of *Political Analysis* focused on the advantages of study pre-registration (Lupia 2008; Monogan 2013; Humphreys, de la Sierra and van der Windt 2013; Anderson 2013).

I propose an addition to the idea of a replication archive that lends credibility in a way similar to study registration, makes all data and model related decisions at all points in time reproducible (if the author so chooses), and serves a pedagogical purpose as well. Simply keep your project (data, tranformation code, model code, and the manuscript) in a Git repository, and post said repository publicly. Git is a distributed revision control system which allows the user to track changes to any file that is text (R and STATA code, LaTeX code, delimited data, etc.), revert to any previous version easily, visualize changes between versions, and a variety of other emminently useful things (creating branches of a project, asynchronous collaboration, etc.). GitHub is an enormously popular web-service that allows the user to host Git repositories publicly (or privately for a price, though they offer free student accounts for 2 years). There are a number of excellent resources for aquainting yourself with Git and GitHub, in particular Pro Git, Try Git, and this excellent answer on StackOverflow. GitHub also has graphical applications available for Mac and Windows machines, as well as integration with Eclipse, Vim, Emacs, and most other text editors with an active community. In addition to the aforementioned offical Git GUIs, there are a number of 3rd party applications that make using Git quite easy (see here for a list). The (justifiably) popular integrated development environment (IDE) for R, R-Studio, also provides integrated support for Git through R-Studio "projects." While STATA does not provide integration with Git, it would be trivial to keep `.do` files in revision control using any one of the above resources. It is worth noting that Git is not the only revision control system (though it is probably the most popular). Subversion and Mercurial are two popular alternatives which could be used for a similar purpose.

A complete research project hosted on GitHub is reproducible and transparent by default in a more comprehensive manner than a typical journal mandated replication archive. With a typical journal replication archive, the final data and code to run the final set of models is provided. This leaves to the imagination most of the details of the data collection and/or

data manipulation that produced the final data set, what model specifications preceded the ones present in the final script, and how the manuscript changed during its journey from idea to publication. With a public Git repository the data, any manipulation code, and the associated models are available at any time that a change was "committed" to a file tracked in said Git repository. Keeping data, data manipulation code, model code, code for visualizations (tables and graphs), along with the manuscript in a Git repository on GitHub (or a similar site such as Bitbucket) thus subsumes and extends the advantages of journal maintained replication archives.

Maintaining your research project on GitHub confers advantages beyond the social desireability of the practice and the the technical benefits of using a revision control system. Making your research publicly accessible in this manner makes it considerably easier to replicate, meaning that, all else equal, more people will build on your work, leading to higher citation counts and impact (Piwowar, Day and Fridsma 2007). Hosting your work on GitHub, because of its popularity in and outside of academia, also increases the probability of your work being seen by people that aren't actively involved in academic political science. Worries about being "scooped" may also be allayed by using a public revision control system, since there is then a public record of your work on the project (as previously noted, you can also keep repositories private).

Additionally, there are pedagogical advantages to this sort of open research. The process by which research ideas are generated, formalized, empirically evaluated, written up, and then (hopefully) published is often opaque to those who have not participated in it. Published papers often seem to have sprung forth from the head of Zeus, absent previous, more imperfect forms. Much of graduate school revolves around learning how to navigate the idea to publication pathway, and all the pitfalls it entails. Greater knowledge of how it is navigated would undoubtedly help in this process. With Git and GitHub, illuminating this would be low-cost.

If open research of this sort was to become a norm in political science, it is hard to imagine that the field would not advance more quickly. Using Git and Github confers non-trivial technical advantages, has a low startup cost given the array of modern software that interfaces with Git, is desireable from a social perspective and an individual perspective, and provides a helpful pedagogical service as well. Although adoption across the field is unlikely (or at least will be a long time in coming), political methodologists are the ideal group of people to be leaders in pushing for transparent, reproducible research, in political science and in related disciplines.

References

Anderson, R.G. 2013. "Registration and Replication: A Comment." *Political Analysis* 21(1):38–39.

Andreoli-Versbach, Patrick and Frank Mueller-Langer. 2013. "Open Access to Data: An Ideal Professed but not Practised." *RatSWD Working Paper Series* 215:1–10.

Humphreys, M., R.S. de la Sierra and P. van der Windt. 2013. "Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration." *Political Analysis* 21(1):1–20.

Lupia, A. 2008. "Procedural transparency and the credibility of election surveys." *Electoral Studies* 27(4):732–739.

Monogan, J.E. 2013. "A case for registering studies of political outcomes: An application in the 2010 House elections." *Political Analysis* 21(1):21–37.

Piwowar, Heather A, Roger S Day and Douglas B Fridsma. 2007. "Sharing detailed research data is associated with increased citation rate." *PloS ONE* 2(3):e308.

Simmons, J.P., L.D. Nelson and U. Simonsohn. 2011. "False-Positive Psychology Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22(11):1359–1366.