Why aren't people sharing their data and code?

Posted by Andrew on 14 September 2015, 9:33 pm

Joe Mienko writes:

> I made the following post on http://academia.stackexchange.com/ a couple of hours ago.
>
> > It is still relatively uncommon for social scientists to share data or code as a part of the peer review process. I feel that this practice runs contrary to notions of replicability and reproducibility and have a desire to voice opposition to instances in which manuscripts are submitted without data and code. Where, however, is such opposition appropriately expressed? I am specifically curious about whether or not it is appropriate to refuse to review an article in the absence of code or data.
>
> Based on my read of your blog, this seems like something you may be interested in writing about.
>
> The original post is here: http://academia.stackexchange.com/questions/54346/demand-code-and-or-data-as-a-condition-of-peer-review

Here's my reply. I see a few things getting in the way of people sharing data and code:

1. It takes effort to clean up data and code to put them in a format where you can share them. I'm recently engaged in a replication project right now (replicating one of my own recent papers), and it's taken a lot of work to set up a clean replication. So that's a lot of it right now: we're all busy and we're all lazy, and setting up datasets for other people is not generally a high priority (although of course it would be if it were required by top journals or by promotion review committees).

2. The IRB is always getting in the way, making you jump through a bunch of hoops if you want to share any data. Much simpler just to lock up or even throw out the data, then you don't have to worry about the busybodies in the IRB telling you that you're not in compliance in some way.

3. Data collection can be expensive in time, money, and effort, and so you might not want to give away their data until you've squeezed all you can out of it. Sometimes there's direct commercial competition, other times it's just the desire in science to publish a new discovery first.

4. This next one is horrible but it does happen: Somebody criticizes your published work and so now you don't want to share your data because they might go through it and find mistakes in your analysis.

5. You're John Lott and you lost all traces of your data, no survey forms, no computer files, no data at all to be found. So nothing to share.

6. You're Diedrik Stapel or Michael LaCour and you never did the study in the first place. You can't share your data because the data never existed. And you wouldn't want to share the code, as it would just be a record of your cheating.

Filed under Sociology | Permalink

39 Comments

1. *Sebastian* says:
   September 14, 2015 at 10:15 pm

   LaCour did share his data (that was, in part, how he got caught) and Stapels at a minimum did "share" data with graduate students (didn't Hauser do that, too)? So 6. doesn't really seem to apply.

   I'd add various forms of proprietary data and non-disclosure agreements as another reason, arguably the most justified. That's fairly common in economics, e.g., when researchers get access to data under specific conditions, be it from private companies or from state sources that might be worried about anonymity if data is too specific. In all of those cases, though, people could and should still be sharing their code.

   - *Andrew* says:
     September 14, 2015 at 10:27 pm

     Sebastian:

     LaCour shared some files with numbers but they weren't the data he was claiming to have. Those data didn't exist. Sharing the data has to mean sharing the actual data, not just some made-up numbers.

     Regarding the proprietary data, I was including that in item 3 above, but I guess it deserved its own item. And of course there are cases where for privacy reasons the data really shouldn't be shared. The IRB isn't always wrong about that!

     - *Sebastian* says:
       September 14, 2015 at 11:10 pm

       I guess it depends on the perspective: from a bird's eye view, clearly you're right and LaCour didn't share any real data. But from the perspective of researchers who asked him for his data at the time, he did share it. I.e. for the question posed in the linked thread, what LaCour did would "count."

       - *Jaime* says:

September 16, 2015 at 11:00 am

Agreed. It's petty to lump in LaCour. After all, it was the community norm (and requirement by Science) that data be shared, and the sharing of an (ostensibly) unrelated dataset that permitted the failed replication attempt that exposed LaCour's fraud. Even more, because both datasets were posted as part of replication data on open data resources (openicpsr.org and dataverse.harvard.edu respectively, see Brookman, Kalla, and Arnow 2015 for details), LaCour's exposure is really because of data sharing not in spite of it.

- *Rahul* says:
  September 14, 2015 at 10:37 pm

  Is research on proprietary data worth it? If no one can independently vet your conclusions, does research becomes a matter of pure faith?

  What are examples of studies using private, non-disclosable datasets that have significant social impact or scientific value?

  - *Sebastian* says:
    September 14, 2015 at 11:38 pm

    I thought the various Facebook studies were quite interesting and, in spite of all justified criticism, I'm definitely glad they got published, but "significant social impact or scientific value" is of course subjective. I work on labor markets and I've seen people get their hands on enforcement data from Brazil that they only got under non-sharing requirements. I'm also glad they published it.

    Beyond strictly proprietary data, there's semi-public data that is generally publicly accessible but you aren't allowed to share it as a researcher, usually because the agencies are concerned about abuse and de-anonymization. That includes, e.g., the Danish labor market data, which is a complete panel of every job change of every Dane, and it's been used for tons of articles. (IIRC, you actually have to have one Danish author to get that data, but I could be misremembering that).

2. *Rahul* says:
   September 14, 2015 at 10:31 pm

   At the core of the issue is this fundamental dichotomy: We like to think & pretend that academic research is an open, noble, transparent enterprise with the pure goal of furthering knowledge.

   In reality, there's a lot of baser issues of careers, tenure, competition, zero sum games, funding money contests etc.

   Often, the optimal strategy for producing reproducible, transparent research is at odds with the approach needed to fulfill the other goals.

   - *Andrew* says:

September 14, 2015 at 10:44 pm

Rahul:

I agree. But one thing luxury that academic researchers have (at least, outside of fast-moving subfields of biology, CS, etc.) is time. Compared to corporate researchers, we are typically in no immediate hurry to finish projects, so we can spend the time to check our results and make data and code reproducible.

- *Rahul* says:
  September 14, 2015 at 10:54 pm

  Andrew:

  No, I disagree, see that's exactly the point! e.g. A young researcher is in a hurry to add publications to strengthen his case for tenure. A tenured, ambitious guy wants to scoop another group working on the same idea.

  Someone else wants to look good on the job market so makes haste. Another guy wants to get an edge by restricting access to his precious data. Someone else knows of obvious flaws in the analysis but if he corrects them his conclusions won't look so strong.

  Lots of reasons to hurry and not check results and not release code and data.

  - *Andrew* says:
    September 14, 2015 at 11:10 pm

    Rahul:

    Sure, a young researcher is in a hurry, he or she might only have a window of two or three years, say, to publish some high-impact work. But that's much less of a hurry than in some corporate settings, where researchers barely have any time at all to write things up.

    In my interactions with colleagues in the corporate world, I've often found them envying that I can spend as much time as I want on a problem, if I really want to.

    - *hjk* says:
      September 15, 2015 at 3:06 am

      Agree that academics have much more time than corporate folks but, to put on my idealist hat, why is that the comparison? Is that a good null hypothesis to demonstrate a significant difference with?

      Also, most senior academics I talk to say they don't really do research or have time to think anymore.

      - *Andrew* says:

September 15, 2015 at 3:25 am

Hjk:

You're hanging out with the wrong sort of senior academic. What's the point of being a senior academic if you don't do research or have time to think? If that's the case, you might as well get another job.

- *hjk* says:
  September 15, 2015 at 3:53 am

  Well, I agree.

  Unfortunately my university is one of the top 5 in the world and one of my mentors just received one of the highest honours possible in this country.

  Coincidentally, I just resigned my job because of my misplaced idealism over these issues (hence commenting too much on your blog the last few days). Got any jobs?

- *Rahul* says:
  September 15, 2015 at 8:06 am

  Andrew:

  He's hanging out with a very commonly found species of academic.

  - *A. Tasso* says:
    September 16, 2015 at 10:28 am

    I think Andrew is hanging out with a very rare species of academic: tenured, hard-money professors. Even tenured professors at soft-money institutions (eg., schools of public health) need to operate on the grant cycle (eg NIH or NSF), which generally means you _don't_ have time to just sit around and think about stuff. The tenure gives you a certain amount of protection but it's very limited.

3. *James Curley* says:
   September 14, 2015 at 10:39 pm

   Interesting points to consider. As of this year, I am now requiring myself and my lab to add code and raw data to our manuscripts as part of the submission process. For my last paper it probably did add a week of my work time to do this. It will probably do the same again for the current paper. I am still going to do it – because I *think* it's right, but I'm not sure I would demand this of others. I think #3 is something that many people grapple with.

   - *Martha* says:

September 15, 2015 at 12:34 am

"As of this year, I am now requiring myself and my lab to add code and raw data to our manuscripts as part of the submission process."

May your tribe increase.

"For my last paper it probably did add a week of my work time to do this."

That sounds like it's worth it for the potential benefit.

- *A. Tasso* says:
  September 16, 2015 at 10:29 am

  "For my last paper it probably did add a week of my work time to do this."

  You could have written an extra paper with that time.

4. *dl* says:
   September 14, 2015 at 10:50 pm

   The more paranoid among us might worry that a reviewer rejects the paper and helps themselves to the data…

   - *Andrew* says:
     September 14, 2015 at 10:54 pm

     Dl:

     The worst was when I got a paper rejected, and one of the reviewers had some comment like, "This paper isn't new; I've seen it floating around on the web for awhile." Grrrrrr.

     - *Rahul* says:
       September 14, 2015 at 11:06 pm

       Andrew:

       In your case, what's your motivation to publish in a Journal? You don't seem to believe in the peer review system, anyways. Nor in the curating role of Journals.

       Posting it on the web you seem to be getting wide dissemination & open discussion & critiques. Your blog probably has more reach than most Journals you publish in.

       So why do you really go through the hoops of peer reviewed publication?

       - *Andrew* says:
         September 14, 2015 at 11:12 pm

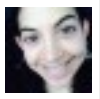
         Rahul:

Hey, I actually covered this in a couple of blog posts last year! See here and here.

5. *Anonymous* says:
   September 14, 2015 at 11:22 pm

   In defense of 4, probably 90% of the code I've seen has errors in it, 75% of those errors don't meaningfully change the results. I think some researchers are concerned that small errors in their code will be blown out of proportion. On balance I'd say code should be published anyways, because long-run we'll start to get to a place where small errors are accepted and large errors exposed. But I can't completely dismiss the concern from some academics that errors will be exaggerated for political purposes, at least in the short term.

6. *Vanessa* says:
   September 14, 2015 at 11:37 pm

   If tools could better capture workflows and data, minimally it would be easier to share a result and allow for replication. The IRB issue is understandable, but in some of these cases it is used as an excuse (as an amendment could always be put into place). 3 and 4 are deeper issues related to the incentive structure for publishing. If success as an academic wasn't so heavily rooted in just that, it would be natural to want to share data and methods to get feedback as soon as they were available.

   - *jrc* says:
     September 15, 2015 at 5:52 pm

     Vanessa,

     Totally agree about better tools. I think one thing that would be really useful is a set of code-writing norms or templates for academics. So for instance, it is crazy to me that I look at people's Stata code and they run 20 regressions and each regression has all of the covariates listed individually. A better way would be for there to be a global macro defined in another .do file ("Macro_Define" or "Do_All" or "Master") that contains the list of covariates that is then referenced in the regression code. This helps the author because if they change anything, they only have to change it in one place and don't accidentally forget to add it to another line. It also makes robustness checks much easier for future replicators – if they want a covariate added, or to use a transformed version of their dependent variable, it is just one variable name added in one place.

     This can be used for other key variables or parameters too – say the level at which fixed-effects are included, or the level at which clustering is defined for standard errors, or even things like number of bootstrap repetitions, or the inclusion/exclusion criteria for the sample.

     I think that if we had norms for organizing and managing analysis files, that would be helpful to everyone: it would make code-writing and organizing faster, it would cut down mistakes, it would allow author's to more easily address referee comments/concerns about covariate choice, and it would make replication and extension that much easier.

It also turns out that for a researcher with very little programming background, coming up with smart ways to do this is hard and takes a lot of work. But I really do think forcing myself to do this has improved the quality of my research – it certainly makes trying out alternative analyses a lot easier (and I hear Andrew yelling "Fork" across a continent, but… some other thread maybe).

7. *Bruce McCullough* says:
September 14, 2015 at 11:50 pm

With respect to Andrew's point (3), withholding the data until the original author has milked it thoroughly,
it is okay to keep a trade secret, but this is research, which depends on building on what others have done.
I argued (McCullough, McGeary and Harrison, J. Money Credit and Banking, 2006):

The objective of an embargo is to permit the author to have sole use of the dataset that he collected. Even without an embargo, some authors will be hesitant to provide data, arguing that it infringes upon the author's competitive advantage. McCullough and Vinod (2003) and Gill and Meier (2000) have noted that such articles cannot be relied upon at least until the embargo period ends, since the accuracy of such rticles cannot be independently verified by other researchers. Not providing data or code, even if due to an embargo, is merely a method by which some journals permit
the author to shift the cost of keeping the dataset to himself (delayed publication) onto the
journal-reading public (in the form of articles whose accuracy cannot be assessed) with
the added expense of retarding scientific progress.

8. *Thomas* says:
September 15, 2015 at 3:03 am

Isn't a lot of this about the perverse incentive structures of modern science, which do not foster truth? If, instead of promoting scientists for "getting results", we promoted them for gathering and preparing useful data sets for everyone to look at it, we'd know much more about the world than we do today. Let the scientists "job" be, not to discover novel truths (which is much rarer than the tabloids suggest), and, rather, to carefully sift through the growing, collective data sets, that have been conscientiously cleaned up for replication by one's peers. We could even award Nobel prizes, both to the people who "came up with the brilliant idea" and to the people whose data made the discovery possible.

(In re 4 in particular), as I usually say, we need to remember that most of science isn't discovering new truths but correcting old falsehoods. So we have to reward the people who make criticism and replication possible, not (so much) the people who stumble on a discovery.

- *Martha* says:
September 15, 2015 at 6:39 pm

+1

9. *Dale* says:
September 15, 2015 at 7:48 am

Thomas – I agree and add #7 (these are not mutually exclusive so I think there is truth in all these reasons):

Academic publication encourages producing multiple studies from the same data – minimal changes in order to get another (and another) publication out of the same data. This is easiest if you are the only one with the data. If we were to reward people for producing the data in usable form rather than just grinding publications out of it, then perhaps it would change.

- *Thomas* says:
September 15, 2015 at 9:25 am

+1

10. *Thomas B* says:
September 15, 2015 at 9:15 am

There are many notable efforts at data sharing out there, for instance, the journal Political Analysis has its Dataverse… http://blog.oup.com/2014/11/replication-data-access-transparency-social-science/

Wharton's Research Data Services is another example: https://wrds-web.wharton.upenn.edu/wrds/

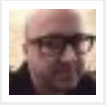
There are hundreds of APIs granting access to otherwise proprietary information. In addition, there are many, many other data repositories of various kinds such as UCI's Machine Learning site — http://archive.ics.uci.edu/ml/
DunnHumby's source file data — http://www.dunnhumby.com/sourcefiles and probably thousands more.

So, let's not fool ourselves into imagining that data isn't being shared.

That said, there are very real barriers to sharing private, "PHI" information: names, addresses, finances, health status, and so on, particularly cross-border and cross-continents. In large part, this has to do with the well-known failure of nearly all encryption systems to protect individual rights to privacy. The recent history of cyber theft is testimony to that.

Shafi Goldwasser, MIT professor of computer science and winner of ACMs Turing Award for her contributions to developing a game theoretic approach to sharing huge amounts of data in gene research. Her approach is generalizable to any area requiring sensitivity in the need for privacy in sharing of information.
http://amturing.acm.org/award_winners/goldwasser_8627889.cfm

11. *Bob Carpenter* says:
September 15, 2015 at 2:31 pm



If (1) is a concern, then why would you trust any results derived from the code in the first place?

- *Andrew* says:
  September 15, 2015 at 2:37 pm

  Bob:

  "Trust" isn't binary. The more I've checked my code and externally validated my claims, the more I trust them. I have indeed published papers with errors. On the other hand, I've published lots of papers which have held up under careful scrutiny, so I think it would be a bit extreme for me to disbelieve every analysis I've done that doesn't have replicable code.

12. *Krzysztof Sakrejda* says:
    September 16, 2015 at 9:06 am

    I disagree with (1), if it was good enough to run the analysis that you're relying on for your conclusions it's good enough to share. Maybe somebody will point and laugh (?), maybe you need to add a warning that the code isn't meant to replicate the analysis exactly (e.g., you fixed some bugs that don't change the final result), but it did it's job and it can do it again (or teach someoene else).

    I already know that most academic code is a mess of spaghetti, but I don't think academic code should be viewed from the same point of view as production code at Airbus or even Facebook. It's ok that it's not neatly done because that's where the field (whichever one) is at.

    The issue that does bother me (also due to a culture of not sharing code or even taking it seriously) is when no reviewer looks at the code, the paper claims to have some generally applicable R package, and the R package is held together with duct tape so it only works for some toy problems (e.g., the package doesn't even bother to combine posterior components on the log scale so it fails once you get past 500 observations).

    - *Bob Carpenter* says:
      September 16, 2015 at 3:58 pm

      I think the main bottleneck isn't a lack of willingness to share, but that researchers tend to use REPL (read, eval, print loop) tools like R and Python and Julia and MATLAB and Stata, so that there's no longer even a record of how they got the results. This was the problem with most of the ARM code we've been trying to update from Andrew and Jennifer's book. So there simply is no longer a (1) that ran the analyses for the paper — the commands used were transient.

      Even when people try to write scripts, the problem is often that they get a data set, write a script to munge it, then modify that script to munge the result, etc., so there's nothing in place to go from the raw data to the analyzed data, even though they think they were scripting it all along. Or tools that evolved over time, each depending on a different R library version for compatibility.

      Another big issue is lack of version control. So the researcher's already modified the tool they used for a paper four years ago and simply don't have a snapshot of what they used.

This is why it's crucial that you need something that runs the analysis end-to-end, from raw data to published results, and you need to tag it in a version control system or freeze it in a tarball for later use. If you do that, it doesn't suffer from this "bit rot" thing Andrew keeps talking about, but doesn't really exist. It was rotten originally — it didn't get rotten over time.

13. *Jaime* says:
September 16, 2015 at 11:12 am

Thinking mostly about code and not data, and about posting rather than share-on-request, in some fields (1) and an associated culture of not sharing leads to a corollary of (3) where those who share are at a disadvantage. Of course, as Andrew points out, journals and funders can change this.

Other tools, like appropriate licenses, might also help nudge community standards. To that end, Matt Might proposed the CRAPL http://matt.might.net/articles/crapl/ From his post (which contains a full copy of the license):

CRAPL-the Community Research and Academic Programming License. The CRAPL is an open source "license" for academics that encourages code-sharing, regardless of how much how much Red Bull and coffee went into its production. (The text of the CRAPL is in the article body.)

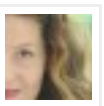
14. *Rahul* says:
September 16, 2015 at 1:54 pm

Andrew: I don't buy the argument: *"It takes effort to clean up data and code to put them in a format where you can share them"*

Why let the perfect be the enemy of the good enough? If the code was good enough to convert your raw data into a publishable paper it is good enough to release.

If you do motivate yourself to clean up the code at a later date, fine; release a new version. But meanwhile posting whatever crappy, ugly code you used online, does more good than harm.

15. *Amy N.* says:
September 16, 2015 at 2:15 pm



Regarding 1), if you are a USA federally funded researcher I hope you are looking forward to the near future where making your federally funded research data publicly accessible is a requirement of the funding. http://scholcomm.columbia.edu/open-access/public-access-mandates-for-federally-funded-research/
This of course influences 2), where more IRBs are becoming sensitive to the requirements mentioned above and are adapting the terms in their suggested consent forms to offer the option of data sharing
You may also be interested in some of the conversations touching on 3) (& #7?) that are working to address the system of incentives/rewards so that openness in publications and data can be rewarded, such as http://osinitiative.org/
Finally, if you are looking for evidence of people who are sharing data, please look at the registry for research of research repositories for data of all types http://www.re3data.org/, which will quickly make you appreciate the first point, that

sharing useful data (FAIR principles
https://www.force11.org/group/fairgroup/fairprinciples) is hard!

16. *Jim Bassingthwaighte* says:
September 16, 2015 at 8:55 pm

While it's true that there are people who are reluctant to share data and models, the expectations are changing rapidly because of the much larger number of people who recognize that "publications" without data and revealed methods of analysis are nothing but advertising and not worth reading. (I paraphrase Donoho at Stanford.) I had the experience a couple of weeks ago, of receiving a model paper to review for a journal (a good one), and while a lot of the equations were given, there was no code. I complained to the editor that I really couldn't review it properly without the code. The next day I received from the editor the coe that he had requested from the authors. I found it excellent, recommended acceptance, with the code in the public domain. People can then build on this code to advance the science further.
Our lab has made code and data available for all papers since the mid 1980s. The early works were in Federal repositories. Most of it, and all the later work is available for inspection and free download at http://www.physiome.org. The payoff is that people use the models and data; that's useful to us in citations, and helps reviewers of our grants and papers.
Before long this kind of responsible, reproducible distribution will be required by NIH and NSF. They currently request it and expect it. Recipients of grant funds should all complete the tasks they have been funded for. Shouldn't that be required?