CS 295B/CS 395B Systems for Knowledge Discovery

Courteous Goodwill:

A Case Study on Reproducible Research via Empirical Software Engineering

The University of Vermont

Themes for today

- Scientific Integrity
- Collegiality and Disagreement
- Power dynamics

Content Warning

We will talk about reproducibility and professional communication today. One topic we will touch on resulted in a student suicide. I do not anticipate this dominating the discussion, but conversations may veer into other sensitive topics.

These topics may be distressing.

If you want to leave now, that is okay.

You can also leave if you just want the time back to study for midterms.

You do not owe me a reason for leaving class.

Why are we talking about these things?

Science is a collaborative, social process

Not just about findings!

- *Doing* science is social
- *Sharing findings* is social

Research methods aren't just about how you investigate research questions

- How we publicize and communicate
- Norms: absorbed vs. taught

Ask vs. Guess culture

This is a classic case of Ask Culture meets Guess Culture.

In some families, you grow up with the expectation that it's OK to ask for anything at all, but you gotta realize you might get no for an answer. This is Ask Culture.

In Guess Culture, you avoid putting a request into words unless you're pretty sure the answer will be yes. Guess Culture depends on a tight net of shared expectations. A key skill is putting out delicate feelers. If you do this with enough subtlety, you won't even have to make the request directly; you'll get an offer. Even then, the offer may be genuine or pro forma; it takes yet more skill and delicacy to discern whether you should accept.

All kinds of problems spring up around the edges. If you're a Guess Culture person -- and you obviously are -- then unwelcome requests from Ask Culture people seem presumptuous and out of line, and you're likely to feel angry, uncomfortable, and manipulated.

If you're an Ask Culture person, Guess Culture behavior can seem incomprehensible, inconsistent, and rife with passive aggression.

Obviously she's an Ask and you're a Guess. (I'm a Guess too. Let me tell you, it's great for, say, reading nuanced and subtle novels; not so great for, say, dating and getting raises.)

Thing is, Guess behaviors only work among a subset of other Guess people -- ones who share a fairly specific set of expectations and signalling techniques. The farther you get from your own family and friends and subculture, the more you'll have to embrace Ask behavior. Otherwise you'll spend your life in a cloud of mild outrage at (pace Moomin fans) the Cluelessness of Everyone.

As you read through the responses to this question, you can easily see who the Guess and the Ask commenters are. It's an interesting exercise.

posted by tangerine at 11:38 PM on January 16, 2007 [1865 favorites]

Famous post on Metafilter

Scenario: Not-close childhood friend asks to stay at your place after being denied before. Is this unforgivably rude?

What does this have to do with science and reproducibility?

What is reproducibility and how did we come to care?

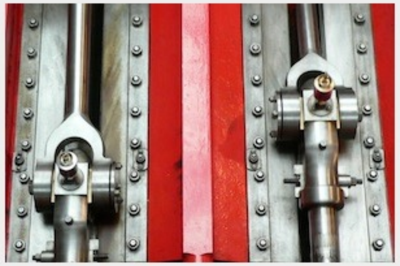
The New York Times Magazine

FEATURE

When the Revolution Came for Amy Cuddy

As a young social psychologist, she played by the rules and won big: an influential study, a viral TED talk, a prestigious job at Harvard. Then, suddenly, the rules changed.

About Artifact Evaluation



In 2011, ESEC/FSE initiated a novel experiment for a major software conference: giving authors the opportunity to submit for evaluation any artifacts that accompany their papers. A similar experiment has since run successfully for several more conferences. This document describes the goals and general mechanics of this process.

If you're just looking for the packaging guidelines, go directly to them.

The rest of this document contains general guidelines about artifact evaluation.
*Individual conferences are **welcome and encouraged** to copy this prose to explain the goals, process, and design to their communities.*

To make things clear to conferences:

Background

A paper consists of a constellation of artifacts that extend beyond the document itself: software, proofs, models, test suites, benchmarks, and so on. In some cases, the quality of these artifacts is as important as that of the document itself, yet our conferences offer no formal means to submit and evaluate anything but the paper. We are creating an Artifact Evaluation Committee (AEC) to remedy this situation.

Goals

Our goal is two-fold: to both reward and probe. Our primary goal is to reward authors who take the trouble to create useful artifacts beyond the paper. Sometimes the software tools that accompany the paper take years to build; in many such cases, authors who go to this trouble should be rewarded for setting high standards and creating systems that others in the community can build on. Conversely, authors sometimes take liberties in describing the status of their artifacts—claims they would temper if they knew the artifacts are going to be scruitinized. This leads to more accurate reporting.

https://artifact-eval.org/about.html

Terminology

A variety of research communities have embraced the goal of reproducibility in experimental science. Unfortunately, the terminology in use has not been uniform. Because of this we find it necessary to define our terms. The following are inspired by the International Vocabulary for Metrology(VIM); see the Appendix for details.

- Repeatability (Same team, same experimental setup)
  - The measurement can be obtained with stated precision by the same team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same location on multiple trials. For computational experiments, this means that a researcher can reliably repeat her own computation.
- Reproducibility (Different team, different experimental setup)*
  - The measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using the author's own artifacts.
- Replicability (Different team, same experimental setup)*
  - The measurement can be obtained with stated precision by a different team, a different measuring system, in a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using artifacts which they develop completely independently.

ACM definitions

How are these related to the other topics

What happens if the reproduction shows that the findings don't hold?

Need to be sensitive:

Today: Work == brand == identity

We are *both* **scientists** and **marketers**

On Professional Courtesy

There two reasons to be skeptical of the rebuttal. The first is that the CACM people have a track record of making mistakes. We should also expect mistakes in their rebuttal, too. Second, if you've been following along, you might have noticed that the rebuttal has a *very* different tone from the two academic papers. The papers were precise and professional. The rebuttal, on the other hand, looks like this:

Claim 7: Critical statistical issues found during reanalysis of study. Hence, **FSE14** and **CACM17** can't be right.

Answer: This is where **TOPLAS19** is most misleading, on multiple accounts. First, their **reanalysis** is only an RQ1 reanalysis. They did not do a reanalysis of our RQ2-RQ4. They gathered their own data, for the same projects we did. For various reasons they couldn't mine all the projects we did. They also could get more data for some projects than we did. This is discussed in Claim 4., above.

#miss They compared their results to our preliminary results in **FSE14**, instead of our **CACM17** results, which are slightly different, and the latter supersedes the former. The **TOPLAS19** authors were aware that our **CACM17** was the definitive version (e.g., see the Introduction of **TOPLAS19**), yet chose to compare to **FSE14**.

#opinion They correct p-values for multiple hypothesis testing, though whether to correct or not in such a way is a matter of debate[10], especially p-values of coefficients within the same regression model. Still, we recognize that some may argue that a balanced correction like the false discovery rate is appropriate. After

Scientific writing is often criticized as dry. And it probably too much is: even a little bit of style makes a paper much more pleasant to read. But we can also take it too far. Writing this charged is a sign that something is off.

This of course goes ways. Vitek's talk was named "Getting Everything Wrong without Doing Anything Right". It's not clear if that contributed to FSE's response, but I'd imagine so. Based on other actions I don't think it was the sole factor, and it happened in an unofficial channel, while this response was the official one. Nonetheless, both sides contributed here. Part of the reason science favors emotionless writing is that it helps avoid feedback loops like this.

Friday readings: reproducibility

Broad community consensus: reproducibility is good

- Verifying findings is important!
- Discussion is important!

Hillel Wayne essay – discourse on Twitter, not in letters!

How you handle the discussion is also important

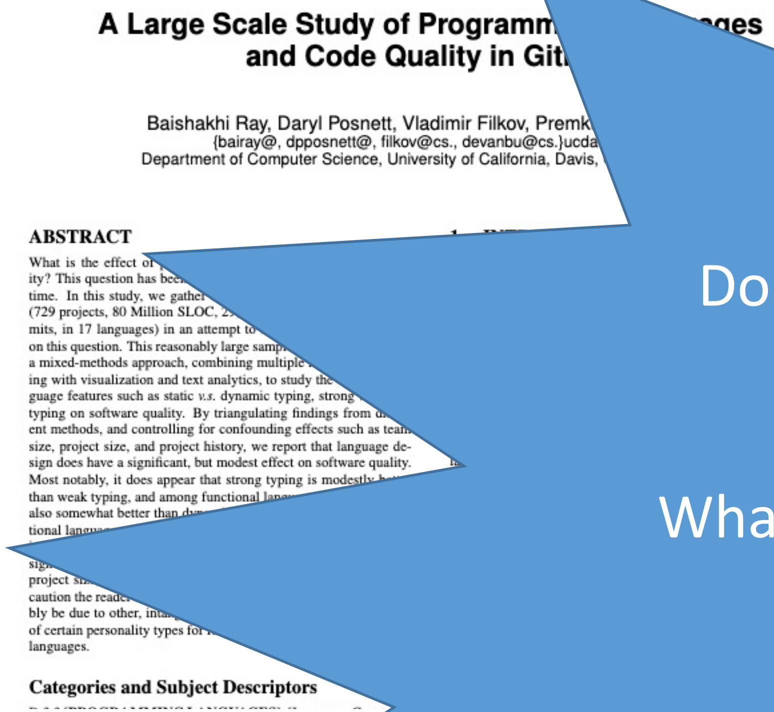
Friday readings: reproducibility

Broad community consensus: reproducibility is good

This of course goes ways. Vitek's talk was named "Getting Everything Wrong without Doing Anything Right". It's not clear if that contributed to FSE's response, but I'd imagine so. Based on other actions I don't think it was the sole factor, and it happened in an unofficial channel, while this response was the official one. Nonetheless, both sides contributed here. Part of the reason science favors emotionless writing is that it helps avoid feedback loops like this.

How you handle the discussion is also important

Power dynamics of parties involved


A Large Scale Study of Programm... ...ages
and Code Quality in Git...

Baishakhi Ray, Daryl Posnett, Vladimir Filkov, Premk...
{bairay@, dpposnett@, filkov@cs., devanbu@cs.}ucda...
Department of Computer Science, University of California, Davis,

ABSTRACT

What is the effect of ... ity? This question has bee... time. In this study, we gathe... (729 projects, 80 Million SLOC, 2... mits, in 17 languages) in an attempt to... on this question. This reasonably large samp... a mixed-methods approach, combining multiple... ing with visualization and text analytics, to study the... guage features such as static v.s. dynamic typing, strong... typing on software quality. By triangulating findings from d... ent methods, and controlling for confounding effects such as team size, project size, and project history, we report that language design does have a significant, but modest effect on software quality. Most notably, it does appear that strong typing is modestly b... than weak typing, and among functional langu... also somewhat better than dyn... tional langu... sig... project si... caution the reade... bly be due to other, inta... of certain personality types fo... languages.

Categories and Subject Descriptors


...mming L... ...de Quality:

21

...t Research

...pecifically, ...ng languages. With ...experimental repe- ...and issues with ...association ...the sta- ...laws ...udy. ...have asserted, ...en task and the number ...ur opinion, even after fixing the ...in to hope for a meaningful

...neering → Soft-

...j, Olga Vitek, and Jan Vitek. 2019. ... ...ction Study. *ACM Trans. Program. Lang. Syst.* 41, 4, ...

NOT saying bad to criticize.
Don't look as an individual problem.

Discussion:
What do we owe the community in our public criticisms?

Original paper ... ...oduc...

First author: Female grad s... ...c... ...st author: very senior male scholar

Bad Behavior & Proposed Resolutions

The network nonsense of Albert-László Barabási

February 10, 2014 in physics, reviews, sophistry | Tags: Albert-László Barabási, Baruch Barzel, DREAM5, Muriel Médard, network, Ofer Biham, partial correlation, regulatory network

In the August 2013 issue of Nature Biotechnology there were two back-to-back methods papers published in the area of network theory:

1. Baruch Barzel & Albert-László Barabási, Network link prediction by global silencing of indirect correlations, *Nature Biotechnology* **31**(8), 2013, p 720–725. doi:10.1038/nbt.2601.
2. Soheil Feizi, Daniel Marbach, Muriel Médard & Manolis Kellis, Network deconvolution as a general method to distinguish direct dependencies in networks, *Nature Biotechnology* **31**(8), 2013, p 726–733. doi:10.1038/nbt.2635.

This post is the first of a trilogy (part2, part3) in which my student Nicolas Bray and I tell the story of these papers and why we took the time to read them and critique them.

We start with the Barzel-Barabási paper that is about the applications of a model proposed by Barzel and his Ph.D. advisor, Ofer Biham (although all last names start with a B, Biham is not to be confused with Barabási):

In order to quantify connectivity in biological networks, Barzel and Biham proposed an experimental perturbation model in the paper Baruch Barzel & Ofer

Critique [edit]

In 2014, Lior Pachter and his student Nicolas Bray published a three-part analy
argued that Barabási has an undeserved reputation for brilliance, because Bar
a small list of examples, in which Barabási's work was subsequently analyzed t

Outside computational biology, critiques have identified various flaws in the me
and the ubiquity of scale-free networks more specifically,[18] his theories on ne
failing to acknowledge the contribution of Derek de Solla Price to the scale-free
version of the Price model, although many properties of the two models do not

https://retractionwa

Retraction Watch

Tracking retractions as a window into the scientific process

PAGES

- How you can support Retraction Watch
- Meet the Retraction Watch staff
  - About Adam Marcus
  - About Ivan Oransky
- Papers that cite Retraction Watch
- Privacy policy
- Retracted coronavirus (COVID-19) papers
- Retraction Watch Database User

Statistical Modeling, Causal Inference, and Social Science

Search

Home Authors Blogs We Read Sponsors

Posted on May 1, 2016 by Andrew

← Previous Next →

No Retractions, Only Corrections: A manifesto.

Under the heading, "Why that Evolution paper should never have been retracted: A reviewer speaks out," biologist Ben Ashby writes:

> *The problems of post-publication peer review have already been highlighted elsewhere, and it certainly isn't rare for a paper to be retracted due to an honest mistake (although most retractions are due to misconduct). Moreover, one could argue that the mistakes in Kokko and Wong's 2007 paper were sufficient to warrant a retraction as they significantly affected the conclusions. But by that logic, a large number of empirical studies should also be retracted due to incorrect statistical analyses or overreliance on fickle p-values, leading to irreproducible results.*

OK, I have no problems so far, except to note that this is never gonna happen.

The part I don't like is what comes next:

> *My concern is that the forced retraction of the original paper sends a bad message to the scientific community. Kokko [co-author of the original*

Discussion time: Less famous, common scenarios

Some actual scenarios (some very common):

- Graduate student feels their work was stolen or they were boxed out
- Graduate student asked to add an author with dubious contributions
- Graduate student finds an error in collaborator’s work
- Graduate student has a scientific disagreement with their advisor
- Graduate student asked to fabricate results

Evidence Puts Doubts on the IEEE/ACM's Investigation

Huixiang Voice Jan 28, 2020 · 5 min read

After the *tragedy that* a Ph. D. candidate Huixiang Chen committed suicide in the University of Florida with a death note claiming that he refused to continue commit acts of academic dishonesty and accused his advisor Dr. Tao Li, IEEE TCCA and ACM SIGARCH launched an ***investigation*** into the alleged reviewing irregularities surrounding the event. We appreciate all the efforts behind this investigation but some evidence from Huixiang Chen's personal laptop put doubts on the result of the investigation.

As the investigation result *claims*:

> *"The committee evaluated whether the paper in question was reviewed according to the established conference guidelines and the review practices of maintaining double blindness, without any contacts from the outside or discussions outside the review process. The committee has determined that there was no evidence of misconduct as part of the paper review process."*

Fabricating results

Unequivocally problematic

Not common, but stakes are very high

Fabricating results: What to do?

Establish outside mentors early in your career

Get help and perspective

Mentor may ask: what makes you say this? Make sure you have evidence and are sure that it’s a fabrication. **Be open to having the wrong read.**

Fabricating results: What to do?

Document exchanges

- “Feeling pressured” is vague; feelings are valid but not actionable
- Get requests in writing and have hard evidence
- Write up summaries after the fact and circulate
- Know your rights (e.g., are you in a 2-party consent state? Can you bring a non-compromised collaborator to meetings as a witness?)

Scientific disagreement

This is normal the more senior you become

Goal of advisors:

train the student to become a peer

What is the nature of the disagreement?

Natural to feel awkward as you transition to a more independent role

Scientific disagreement: What to do?

Malpractice/bad behavior not taken lightly

Try to understand why your advisor might disagree:

- Are they familiar with the methods? Perhaps they feel they cannot advise you on this and don't know how to say that?
- Do they understand the problem? Be assertive! You are becoming the expert!
- Talk to your mentors and collaborators; practice being heard.

Error in collaborator’s work

Researchers are not infallible

- People make mistakes
- Peer review is imperfect
- Unhealthy to hold researchers to impossible standards

People are sensitive to criticism when it’s tied to their integrity and sense of self (most researchers)

This is fundamentally unscientific, but a reality

Error in collaborator’s work: What to do?

Perfect world: **always talk it over with your advisor**!

We live in an imperfect world, so…

If the collaborator is *not* your advisor, then **talk to your advisor**

Your advisor may know the person better, know how to approach (if appropriate)

If the collaborator *is* your advisor, then **talk to your outside mentor**

Does the error cast doubt on the integrity of the work?

Dubious contributions of coauthors

Authorship disputes: extremely common

Authorship is discipline-specific

Advisors should discuss what constitutes authorship early and often (e.g. ACM guidelines)

“Invisible” contributions

- High-level discourse (in every meeting)
- Backchannel conversations

Dubious contributions of coauthors: What to do?

Where do you fall on the author list?

What contributions does that person believe they have made?

Everyone should be able to articulate what they contributed to a publication

DO NOT accuse that person of not making contributions

Can think of **authorship as contract**…

and can go both ways

Stolen work/boxed out

EXTREMELY common sentiment

Contemporaneous discovery

Truly stolen work far less common than feeling/being boxed out

Stolen work hard to prove

Boxed out == social phenomena

Stolen work/boxed out: What to do?

BE GENEROUS with others

- Instinct will be to be defensive.
- Better to trust but verify.

BE GENEROUS with yourself

You have many good ideas.

Sharing your ideas as audition: how do people treat you? Would you work with them in the future?

Antidote is for you to publicize your work and be generous with credit!

What does this have to do with reproducibility?

Theme: Trust

Trusting positive intent of authors

(they want to do good science!)

BUT feel free to be skeptical of results

(this is scientific!)

Reproducibility as a social process

- As a conversation
- Focus on the big ideas