The Reproducibility versus Debuggability of Research

Michael Rabinovich • *Case Western Reserve University*

The computer science community has a long-lasting debate on research reproducibility. At the heart of the issue is the question of whether the data used in a particular study should be publicly available. On one side of the debate, the argument is that without the data, the research results can't be independently verified, which makes the research — no matter how interesting in itself — useless. In particular (the argument goes), this automatically makes any research results without the underlying data unpublishable. The counterargument is that some of the most valuable insights are based on data that comes from industrial sources and can never be released due to privacy, contractual, or proprietary concerns. The consequence of taking an uncompromising position on data availability — and hence research reproducibility — is thus to deprive the research community of the insights the industrial data generates.

This debate flares up from time to time. Different subcommunities arrive at different conclusions, and then these conclusions change with time. The *New York Times* published a good account of these issues a couple of years ago.[1] I personally encountered this controversy most recently on two occasions: during the 2012 program committee meeting for the World Wide Web conference, where a lengthy debate on whether to accept some papers revolved around public availability of the data used; and during the program committee discussions for the 2013 Passive and Active Measurement (PAM) conference, which in its guidelines warned that submissions that didn't make their data available could be rejected without review. The PAM conference warning had a "loophole" asking authors to explicitly explain why they weren't releasing the data. In my view, the warning didn't achieve its goals of encouraging more data release: those who didn't plan a release simply put an obligatory disclaimer.

The resulting uncertainty in what constitutes legitimate research hurts the research environment. It adds more randomness to the already often-unpredictable reviewing process. An otherwise worthwhile paper might be accepted or rejected for publication depending on which side of the debate a reviewer happens to belong to. Thus, it's important to develop a consensus on this issue, at least within individual research fields. My thoughts here are based on my experiences in the Internet and networking areas.

Is Research Based on Unreleased Data Useless?

I mentioned earlier that cleansing published research of studies based on unreleased data would deprive the community of valuable insights. But without reproducibility, are these insights really valuable? The answer, at least from the networking community I'm mostly familiar with, appears to be yes. A quick look at the 2013 program for the Internet Measurement Conference (the premier conference on the topic) shows that among 11 papers presented on the first day, four were based on industrial data that seem unlikely to be released. In general, it's well understood among the networking research community that releasing datasets increases the work's impact and makes it more valuable; yet it's also understood — as evidenced by papers without released datasets being routinely accepted — that not doing so doesn't render the contributions worthless. I personally agree with this position. These contributions add value in several ways. First, they typically include novel methodologies that are reproducible and advance the state of the art. Second, the results — even if not reproducible — raise questions and at least put out a conjecture about the issues at hand. Third, experimental computer scientists, when they design their technologies, need basic understanding about

Published by the IEEE Computer Society 1089-7801/14/$31.00 © 2014 IEEE

their target environment's behavior to drive their reasoning. If results from a well-conceived study with a convincing methodology help form such understanding, most people would prefer to have it even if it isn't fully verified. (This, of course, isn't black and white and depends on the consequences of being wrong. Also, if the context formed by a faulty study turns out to be wrong, the designs built on it might also be misguided and will need to be rolled back. Thus, this point has two sides to it.) Finally, even if other researchers can't verify the specific results without the data, they can often confirm those results using different, independently obtained data. The networking community values such independent confirmations highly. Conferences, notably PAM, explicitly solicit contributions containing confirmation of previous studies. In fact, these confirmations are often needed regardless of whether the original study released the dataset. Results — even when derived faithfully from the released data — might not stand up to subsequent scrutiny for many reasons; these range from faulty data collection instrumentation to data being too narrow to be representative of the phenomenon at hand.

Thus, in networking research, whether or not the underlying data is released, research results must be confirmed by independent studies on independent data. Only on accumulating a body of confirming independent evidence do the results enter into the collective mindset that forms the networking research foundation. This process is gradual, with no clear threshold. But studies based on both released and unreleased data follow it.

We can conclude that equating the availability of underlying data with research reproducibility implies a narrow meaning of reproducibility — specifically the ability to verify that the research results were faithfully derived from the data at hand. Although useful, a broader notion of reproducibility involves confirming the results across different data sources; whether the underlying data is released doesn't affect research reproducibility in this broader sense. In fact, I would say that most value in data release comes not from the ability to verify the research that used the data first, but in facilitating further research.

If Not Reproducibility, Perhaps Debuggability?

I've argued that reproducibility in its narrow sense is often unattainable and shouldn't necessarily be a prerequisite for published research in computer science. However, mistakes in data analysis do happen, and it seems important that, if questions arise, the authors themselves should at least be able to go back to their data and debug their work. For this, they must keep their data long after the research is complete. But datasets are becoming increasingly large — hundreds of gigabytes are common, and terabytes

aren't unheard of. An experimental scientist accumulates these datasets quickly, and if he or she can't discard them for years, how can the scientist ensure their availability?

Universities commonly offer a central data archiving service, but this involves recurring costs. Such costs are problematic for an academic researcher when they extend beyond the project's life because most funding is earmarked to ongoing projects. Public clouds offer inexpensive storage (Amazon Glacier being designed particularly for archival needs), but they again require recurring costs. Several efforts are ongoing to operate cooperative data storage targeting specific research areas, such as the Encode project for bioinformatics (www.genome.gov/10005107) and the Inter-University Consortium for Political and Social Research for social sciences (www.icpsr.umich.edu/icpsrweb/landing.jsp). Other consortia target general research and engineering data (for instance, the DataNet Federation Consortium; http://datafed.org) and nonresearch digital data artifacts (as with the National Digital Stewardship Alliance; www.digitalpreservation.gov/ndsa/about.html).

However, the common theme of all these efforts is to make the data available, which wouldn't help with unreleased data. Even when a research archive offers a complete embargo on data access, as does the Netherland's Data Archiving and Networked Services (www.dans.knaw.nl/en/content/data-archive), it requires that the data be submitted in a readable format and insists that "any privacy-sensitive information must be deleted from the data," as stated in the FAQ document linked from www.dans.knaw.nl/en/content/data-archive/depositing-data, making it unsuitable for unreleased datasets.

What's missing to facilitate long-term research debuggability is a free archival service that, rather than storing released data for sharing, would merely safe-keep the data for authors, who would maintain sole ownership and control. In fact, the authors should be able to submit encrypted datasets and keep the keys so that no one can access the data but themselves (the keys are small, so keeping them is much less daunting for a researcher than is keeping the data itself). Because this service must be free, it would most logically be run by the government (perhaps by the US National Archives and Records Administration or a similar agency), a national lab, or a university under a government contract. Generally, storage is cheap enough now that this type of a universal archive might be feasible, especially given that online access isn't required (so something such as tapes stored on shelves would work fine). The system could be operated professionally and efficiently, rather than in the myriad ad hoc ways individual researchers do it today. In fact, the government would probably save money because it would remove the need for individual researchers to maintain these inefficient archival facilities, which usually are funded from government grants anyway. Once such an archive is operational, submitting encrypted datasets that ensure research debuggability could become a requirement for paper acceptance at more rigorous venues and a component in data management plans in grant proposals.

In an ideal world, we would make all research data available to everyone. The reality is that this ideal is unattainable, but this shouldn't preclude a middle-ground solution that, while short of the ideal, would be a huge step toward facilitating sound research.

Reference

1. J. Markoff, "Troves of Personal Data, Forbidden to Researchers," *New York Times*, 21 May 2012, page D1.