Making Social Media Research Reproducible

Luke Hutton and **Tristan Henderson**
School of Computer Science
University of St Andrews
St Andrews, Fife KY16 9SX, UK
{lh49,tnhh}@st-andrews.ac.uk

Abstract

The huge numbers of people using social media makes online social networks an attractive source of data for researchers. But in order for the resultant huge numbers of research publications that involve social media to be credible and trusted, their methodologies, considerations of data handling and sensitivity, analysis, and so forth must be appropriately documented. We believe that one way to improve standards and practices in social media research is to encourage such research to be made *reproducible*, that is, to have sufficient documentation and sharing of research to allow others to either replicate or build on research results. Enabling this fundamental part of the scientific method will benefit the entire social media ecosystem, from the researchers who use data, to the people that benefit from the outcomes of research.

Introduction

The use of social media and online social networks (OSNs) such as Facebook and Twitter for research has exploded in recent years, as researchers take advantage of access to the hundreds of millions of users of these sites to understand social dynamics, health, mobility, psychology and more. But while such research can be of high value, the myriad sites from which data can be collected, the types of data available, the ethical concerns, the various data collection techniques and analysis techniques means that there is huge variation in methodology, all of which must be documented if people are to understand and build on such research. Indeed such documentation is a fundamental part of the scientific method; if other people are to replicate and reproduce research, that research and relevant outputs must be documented and shared to enable this. Even if people do not intend to conduct further research, sharing and documentation will help make research results more credible. In short, we believe that good social media research should be *reproducible* research.

In this paper we draw on a recent study that examined the state of reproducibility in social media research, specifically focusing on practices in three areas: *code*, *method* and *data*. We describe the current state of the art, and propose recommendations for these areas. We then outline some overall challenges that we believe the community is well-timed to tackle.

Why make research reproducible?

Reproducibility and the ability to replicate and build on the research of others has long been recognised as a fundamental part of the scientific method and necessary for scientific publications (Lubin 1957). Unfortunately, many studies have shown that much research is not reproducible (Bonnet et al. 2011; Hornbaek et al. 2014; Ioannidis et al. 2009). This worrying situation means that much of science is not credible, in that many results cannot be verified, which has led to calls for boards to verify, replicate and certify experiments (Baker 2012). Calls for better reproducibility in research have come from researchers in fields including geoscience, bioinformatics, statistics, physics and computer science (Yale Law School Roundtable on Data and Code Sharing 2010), while funding agencies are increasingly mandating that data be shared (National Science Foundation 2013) and particularly in the case of biomedical research, made reproducible (Collins and Tabak 2014).

Calls for the latter, for better reproducibility in computational sciences (e.g., Donoho et al. (2009)), are particularly relevant to social media research. Given that most social media are delivered online, it is unlikely that there are many social media research studies that do not involve the use of computation, be it for data collection, data processing, or data analysis. Donoho et al. argue that computation is both indispensable for modern research, but moreover forms a third scientific methodological branch distinct from deductive and empirical methodologies. And yet, they say that "computational science has nothing like the elaborate mechanisms of formal proof in mathematics or meta-analysis in empirical science". Building mechanisms to enable the reproducibility of computational work will thus help improve the science itself.

The state of the art

In a recent study (Hutton and Henderson 2015b) we searched various research paper databases to find 901 papers published between 2011 and 2013 in a variety of 26 computer science, psychology, anthropology and communications venues that included in their abstracts the terms social network, online social network, or the names of various

popular OSNs. From the original 901, we found that 505 papers actually used data from OSNs.[1] We then examined each of these papers to determine how "reproducible" they were.

Following Thompson and Burnett (2012), we characterise reproducibility by three elements:

- does a research artefact share the source code, tools, and workflows required to replicate the data collection that led to the creation of the artefact?
- does an artefact encode the scripts and methods that conduct analyses and produce the components of an artefact?
- does an artefact include the dissemination of raw data and other resources required for replication?

We can broadly describe these as the documentation and dissemination of code, methods and data respectively. Each of the 505 papers that we examined were tested against ten criteria which addressed these categories:

1. *Code:* Were source code or software tools shared?
2. *Method:* Was the source OSN used described?
3. *Method:* Was the measurement or sampling strategy used to collect data described?
4. *Method:* Was the length of the study described?
5. *Method:* Were the number of participants or data points described?
6. *Method:* Were mechanisms for processing data, sanitising, aggregating, anonymising and so forth described?
7. *Method:* Was consent obtained from participants in the OSN, and if so was this described and how?
8. *Method:* Were participants briefed or debriefed to explain the research, and was this described?
9. *Method:* Were there any ethical concerns and how were these addressed? Was IRB (Institutional Review Board) or ethics committee approval needed and obtained?
10. *Data:* Were datasets used for the research either cited (where other data were used), or provided (where original data were collected)?

Unfortunately, like the aforementioned surveys in other disciplines, we found that many of these papers were not reproducible. Indeed, we only found one paper that met all ten criteria. Figure 1 highlights how well all criteria were reported across the survey.

Reproducible code?

Code and protocols were not particularly well documented. Computer science papers were better at this, but social science papers rarely share code and protocols. None of the psychology papers and only 5.6% of communications papers we examined shared code. This may be due to conventions in these fields, and so one open challenge is to open up dialogue or sharing of best practices across communities.

[1]Full details are available in the paper and the dataset of papers studied, which can be found in the FigShare data archive (Henderson and Hutton 2014).

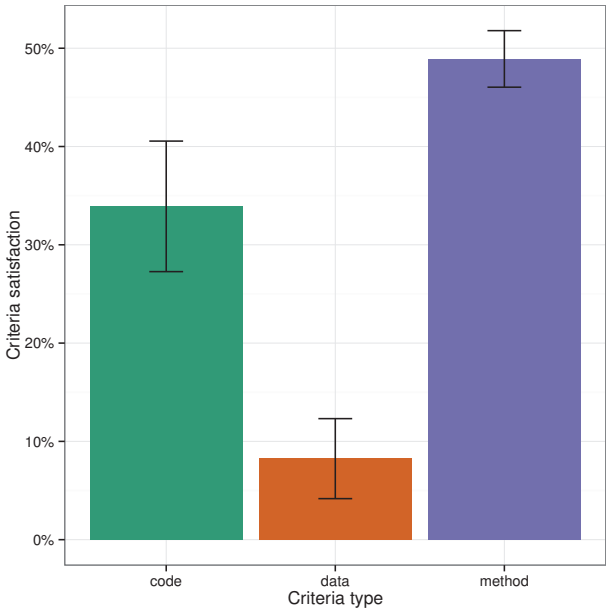

Figure 1: Barplot showing how well our three criteria types are met by the papers we study, with standard error between disciplines shown. Data-sharing is particularly poor across most disciplines, while reporting of methodologies is generally stronger. The sharing of code is the most variable practice between disciplines.

Approximately a quarter of workshop and journal papers share their code, while 41.4% of conference papers do so. We can attribute this to researchers perhaps being reticent to share unfinished code with a workshop audience, while more computer science conferences are requiring authors to provide supplementary materials, such as source code, with their papers.

There exist many new initiatives for capturing and archiving code for research experiments, for instance recomputation.org (Gent 2013), the Software Sustainability Institute (Crouch et al. 2013), or RunMyCode (Hurlin, Pérignon, and Stodden 2014), which can all help with the sharing of code, tools, or even virtual machines that encapsulate all of the code and data needed to run the analysis to create plots, or an entire research paper. Researchers can also use source code hosting services such as GitHub,[2] but such services do not provide specific support for maintaining research code.

Reproducible method?

Of the three areas that we examined, the sharing of methodologies was by far the strongest. Basic parameters such as the source OSN, sampling strategies, length of studies and the number of participants were almost (but not always) universally reported. But other parameters to do with the handling of data and participants, and in particular the reporting of ethical concerns, were generally not discussed. The

[2]GitHub: http://www.github.com

venues that were best at reporting these were security and privacy venues; this might be due to the fact that venues such as SOUPS (Symposium on Usable Privacy and Security) requires the describing of ethical concerns and how these were tackled, while WPES (Workshop on Privacy in the Electronic Society) allows appendices with supplementary information to be included. We note that ICWSM and other venues, such as ACM SIGCOMM, have started asking submitters whether ethics approval was obtained. But to aid the reproducibility of work, and to allow readers to truly understand how a study was conducted, it would be good to consider adopting best practices from SOUPS and WPES, and allowing additional appendices. Another possible route for improving standards might be to look at some of the proposed guidelines for social media research (McKee 2013; Rivers and Lewis 2014), although some of these are quite specific to subsets of such research and may not be appropriate to social media research at large.

Reproducible data?

Only 6.1% of the papers in our survey shared or cited data. This was somewhat surprising given moves by journals and funding bodies to encourage or even mandate the sharing of research data. Data sharing is one area where ICWSM does relatively well, with the existing ICWSM data sharing initiative[3] meaning that 35.4% of the (admittedly small) number of papers that shared data came from ICWSM. But overall the picture is poor, and calls to improve this have already been made (Huberman 2012).

Data sharing and data reuse in itself may not be sufficient for reproducibility. In separate work one of the authors runs the CRAWDAD wireless network data archive, which provides over 120 wireless datasets to 10,000 researchers, who have used data for research papers, education, standards development and more (Kotz and Henderson 2005). The sharing of these datasets has clearly been beneficial to all of these researchers, but merely offering datasets for download does not guarantee that readers will be able to find the datasets used in a paper. We need to enable mechanisms for appropriate data citation (Henderson and Kotz 2015) to allow readers of papers to understand what data have been used, allow persistent identifiers for citation, and to provide credit and attribution to authors. The Force 11 Data Citation Principles[4] go some way towards codifying these, and could perhaps be considered by the community for social media datasets.

General education about how best to work with and use data might also be useful for improving standards; groups like Data Carpentry[5] have experience in this space.

Reproducing workflows

From looking at the state of the art it is clear that improvements could be made to further the reproducibility of social media research, through sharing and documentation of code, method and data. One overall way to do this might be to build mechanisms for capturing experimental workflows and exporting these for distribution and replication are needed. The need for easy automated capture of experimental context and workflow has already been noted as a key driver for improving reproducibility (Davison 2012), and the e-science community in particular has developed various workflow tools (De Roure et al. 2010).

Capturing the workflow of social media experiments has particular challenges, and goes beyond sharing the source code of tools used to collect and analyse data. We suggest that to support reproducibility, the following aspects of the workflow (code and method) must be captured:

1. **Data sampling strategy** First, we must consider about whom the data we collect relates. Is it possible to encode details about the population sampled? This might mean users who tweeted using a certain hashtag, or a stratified distribution of participants meeting certain demographic criteria. Motivated by our survey, in which these details were reported in 71.9% of cases, it is important to be able to encode these details such that others can replicate the procedures and compare their findings. This introduces particular issues where the content collected is unique to a point in time, such as Denef et al.'s study of tweeting practices during the 2011 London riots (Denef, Bayerl, and Kaptein 2013). Even if we can encode these sampling details, it may not be possible to collect a similar distribution of content in a replication, as the APIs provided by social media services do not usually provide archival access, and streaming APIs provide a limited subset of all published content (Morstatter et al. 2013).
2. **Ethical procedures** We argue that the ethical procedures of an experiment should be encoded to support replications, which as we noted earlier, does not often happen. Were the people whose data were collected considered human subjects, and as such were they informed about the procedures of the experiment, and did they provide consent to participate? For all experiments, the steps taken to preserve the privacy of participants should be encoded, such as which personally identifiable information was removed, and whether sensitive data such as locations were obfuscated. All ethical procedures, including briefing documentation, consent forms, and sanitisation processes, are an important part of the workflow as they have implications for the analysis of a study. For example, when replicating previous analyses, understanding how data have been sanitised is necessary to put any findings in context, and avoid making assumptions the sanitised data do not support. As participants in social media experiments are willing to share different types of data depending on the purpose of the research (McNeilly, Hutton, and Henderson 2013), documenting the consent and briefing materials given to participants is necessary to conduct replications in the same context, and to better understand how the results were produced.
3. **Data collection** Encoding the workflow of the data collection itself is essential. To an extent, this is encoded in the source code for tools used to conduct experiments, but replicating this binds other researchers to the same

[3] http://icwsm.org/2015/datasets/datasets

[4] https://www.force11.org/datacitation

[5] http://datacarpentry.org/

platforms and environments of the original research, and may demand expertise in these environments to enable replications. We argue that the data collection procedures of an experiment should be encoded in a form which is portable, human-readable, but can also be parsed by software to support the conduct of a study. Enumerating the services data are collected from, the types of data collected and how they are sanitised, provides a specification of the requirements of a study. In replications, this can support applications for ethical approval by easily communicating the requirements of a study, and makes it easier for others to conduct a study under the same conditions.

4. **Analysis** Finally, we should capture how research outputs are produced as a result of the data collection process. Understanding what pre-processing was performed on a dataset collected from social media and which analyses were conducted is important to validate the findings of others, and to directly compare results between studies. As discussed earlier, this is a challenge for all computational research, and is not unique to social media studies.

Considering these four workflow steps, we are developing tools for social media experiments (Hutton and Henderson 2013) to support the encoding and sharing of social media experiment workflows. Researchers can specify a policy that governs data collection, sanitisation, processing and captures ethical concerns and can generate consent forms for participants. These policies can then be shared to allow other researchers to replicate experiments. We are working towards supporting all four requirements we outlined.

Our tools' policy language allows the full data collection procedures to be encoded, including how data are sanitised, allowing other researchers to see at what stage in an experiment data are collected, and how they are sanitised and stored. For example, in an experiment where location data were collected, our policy can show whether the data were collected at full resolution for analysis, then coarsened to reduce the chance of the participant being reidentified from a resulting dataset. As our policy language enumerates all data collection procedures, we can generate consent forms which allow participants to understand exactly how their data are used, rather than relying on boilerplate forms alone. These policies are human-readable so that other researchers can understand the procedures of an experiment, even if they do not use our tools. If they do, however, our framework can enforce the policy to ensure an experiment doesn't collect data beyond its remit. We have therefore made significant progress towards the second and third steps we identified, but open problems remain.

Sharing workflows will allow researchers to reproduce code and method, but data must also be shared. Thus we are interested in how we can extend support into the rest of the workflow, allowing sampling strategies to be encoded and shared, and whether we can support the sharing of other research artefacts. One such approach might be to embed an instance of our tools when sharing other items such as source code and raw data, allowing policies to be re-used, and provide programmatic access to social media datasets in the future, without the need for a third party service to be available.

So far, we have used this system for our own social media experiments, investigating location-sharing services (Hutton, Henderson, and Kapadia 2014), and informed consent in Facebook studies (Hutton and Henderson 2015a), and to replicate previous experiments from the social media privacy literature (Hutton and Henderson 2015b). Our tools do not yet tackle all of the challenges we have identified, and as it is not necessarily the case that these experiments are representative of all OSN research, we would welcome other researchers to use our system. We are also keen to discuss whether our model of the social media experiment workflow is appropriate to all such experiments, and what extensions we could develop to benefit researchers in the community. We would like to contribute to the development of best practices for research in this area, building on the challenges we have identified.

Further challenges

Beyond documenting and sharing code, method and data, some additional overall challenges must also be addressed. One huge challenge with introducing new systems for data capture, or mechanisms for sharing data, is *sustaining* these systems. Running services that allow researchers to run or replicate experiments would require community support and funding beyond the scale of the current ICWSM data sharing service.

Another particular problem with capturing social media experiments, versus experiments in some other branches of computational science, is the interactions with both people and *other systems* that may be outside of a researcher's control, e.g., the OSNs. For instance, when we attempted to replicate the one reproducible paper from our survey, we ran into trouble because the Facebook APIs had changed since the original researchers' experiment, and we were unable to retrieve the same information. Deciding how to deal with such changes is difficult, but better documentation is a good start. Some fifteen years ago, Floyd and Paxson (2001) recognised the difficulties with simulating the Internet. Perhaps we now need to turn our efforts to recognising and addressing the challenges of capturing the relevant parts of the Internet that constitute a research artefact and making such artefacts reproducible.

Perhaps the biggest challenge is that of creating appropriate *incentives* for researchers to make their research more reproducible. After all, any of the suggestions proposed here involve additional effort on the part of busy and overworked researchers. Our experiences with data sharing indicate that "sticks" such as government or funding body mandates do not necessarily seem to work, as we have not seen a significant increase in dataset contributions despite mandates from the NSF in the USA, or the Research Councils in the UK. Perhaps we should instead investigate "carrots". There is evidence that data sharing, for instance, will increase the impact of research (Piwowar, Day, and Fridsma 2007). We have also had some success with changing culture through publication habits. For instance the ACM Internet Measurement Conference requires that data be shared, not for all pa-

pers, but if a paper is to be considered for the Best Paper Award. Such steps can encourage researchers to make their work more reproducible without being overly punitive.

But perhaps the best incentive is to return to Donoho et al. (2009), who say that "striving for reproducibility imposes a discipline that leads to better work."

Conclusions

We have posited that one way to develop better standards and practices in social media research is to encourage such research to be reproducible. By documenting and sharing research in such a way that others can replicate or build on results, the original researchers will hopefully improve their practices, while other researchers will benefit by improved access to research outputs, but also by the sharing and dissemination of research practices themselves. We have offered a brief look at the state of the art in reproducibility, outlined our work-in-progress workflow tool, and touched on some (but not all) of the outstanding challenges that need to be addressed.

References

Baker, M. 2012. Independent labs to verify high-profile papers. *Nature*. Available online at http://doi.org/10.1038/nature.2012.11176.

Bonnet, P.; Manegold, S.; Bjørling, M.; Cao, W.; Gonzalez, J.; Granados, J.; Hall, N.; Idreos, S.; Ivanova, M.; Johnson, R.; Koop, D.; Kraska, T.; Müller, R.; Olteanu, D.; Papotti, P.; Reilly, C.; Tsirogiannis, D.; Yu, C.; Freire, J.; and Shasha, D. 2011. Repeatability and workability evaluation of SIGMOD 2011. *ACM SIGMOD Record* 40(2):45–48.

Collins, F. S., and Tabak, L. A. 2014. Policy: NIH plans to enhance reproducibility. *Nature* 505(7485):612–613.

Crouch, S.; Hong, N. C.; Hettrick, S.; Jackson, M.; Pawlik, A.; Sufi, S.; Carr, L.; De Roure, D.; Goble, C.; and Parsons, M. 2013. The Software Sustainability Institute: Changing research software attitudes and practices. *Computing in Science & Engineering* 15(6):74–80.

Davison, A. 2012. Automated capture of experiment context for easier reproducibility in computational research. *Computing in Science & Engineering* 14(4):48–56.

De Roure, D.; Goble, C.; Aleksejevs, S.; Bechhofer, S.; Bhagat, J.; Cruickshank, D.; Fisher, P.; Kollara, N.; Michaelides, D.; Missier, P.; Newman, D.; Ramsden, M.; Roos, M.; Wolstencroft, K.; Zaluska, E.; and Zhao, J. 2010. The evolution of myExperiment. In *Proceedings of the IEEE 6th International Conference on E Science (e-Science)*, 153–160.

Denef, S.; Bayerl, P. S.; and Kaptein, N. A. 2013. Social Media and the Police: Tweeting Practices of British Police Forces During the August 2011 Riots. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, 3471–3480. New York, NY, USA: ACM.

Donoho, D. L.; Maleki, A.; Rahman, I. U.; Shahram, M.; and Stodden, V. 2009. Reproducible research in computational harmonic analysis. *Computing in Science & Engineering* 11(1):8–18.

Floyd, S., and Paxson, V. 2001. Difficulties in simulating the Internet. *IEEE/ACM Transactions on Networking* 9(4):392–403.

Gent, I. P. 2013. The recomputation manifesto. Available online at http://arxiv.org/abs/1304.3674.

Henderson, T., and Hutton, L. 2014. Data for the paper "Towards reproducibility in online social network research". Available online at http://doi.org/10.6084/m9.figshare.1153740.

Henderson, T., and Kotz, D. 2015. Data citation practices in the CRAWDAD wireless network data archive. *D-Lib Magazine* 21(1/2).

Hornbaek, K.; Sander, S. S.; Avila, J. A. B.; and Simonsen, J. G. 2014. Is once enough?: On the extent and content of replications in human-computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3523–3532.

Huberman, B. A. 2012. Sociology of science: Big data deserve a bigger audience. *Nature* 482(7385):308.

Hurlin, C.; Pérignon, C.; and Stodden, V. 2014. RunMyCode.org: A research-reproducibility tool for computational sciences. In Stodden, V.; Leisch, F.; and Peng, R. D., eds., *Implementing Reproducible Research*. Chapman and Hall/CRC.

Hutton, L., and Henderson, T. 2013. An architecture for ethical and privacy-sensitive social network experiments. *ACM SIGMETRICS Performance Evaluation Review* 40(4):90–95.

Hutton, L., and Henderson, T. 2015a. "I didn't sign up for this!": Informed consent in social network research. In *Proceedings of the 9th International AAAI Conference on Web and Social Media (ICWSM)*.

Hutton, L., and Henderson, T. 2015b. Towards reproducibility in online social network research. *IEEE Transactions on Emerging Topics in Computing*. To appear.

Hutton, L.; Henderson, T.; and Kapadia, A. 2014. "Here I am, now pay me!": Privacy concerns in incentivised location-sharing systems. In *Proceedings of the 2014 ACM Conference on Security and Privacy in Wireless & Mobile Networks*, 81–86.

Ioannidis, J. P. A.; Allison, D. B.; Ball, C. A.; Coulibaly, I.; Cui, X.; Culhane, A. C.; Falchi, M.; Furlanello, C.; Game, L.; Jurman, G.; Mangion, J.; Mehta, T.; Nitzberg, M.; Page, G. P.; Petretto, E.; and van Noort, V. 2009. Repeatability of published microarray gene expression analyses. *Nature Genetics* 41(2):149–155.

Kotz, D., and Henderson, T. 2005. CRAWDAD: A Community Resource for Archiving Wireless Data at Dartmouth. *IEEE Pervasive Computing* 4(4):12–14.

Lubin, A. 1957. Replicability as a publication criterion. *American Psychologist* 12(8):519–520.

McKee, R. 2013. Ethical issues in using social media for health and health care research. *Health Policy*.

McNeilly, S.; Hutton, L.; and Henderson, T. 2013. Understanding ethical concerns in social media privacy studies.

In *Proceedings of the ACM CSCW Workshop on Measuring Networked Social Privacy: Qualitative & Quantitative Approaches*.

Morstatter, F.; Pfeffer, J.; Liu, H.; and Carley, K. M. 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *Proceedings of the 7th International AAAI Conference on Web and Social Media (ICWSM)*.

National Science Foundation. 2013. Other Post Award Requirements and Considerations. Available online at http://www.nsf.gov/pubs/policydocs/pappguide/nsf13001/aag_6.jsp#VID4.

Piwowar, H. A.; Day, R. S.; and Fridsma, D. B. 2007. Sharing detailed research data is associated with increased citation rate. *PloS one* 2(3):e308+.

Rivers, C. M., and Lewis, B. L. 2014. Ethical research standards in a world of big data. *F1000Research*.

Thompson, P. A., and Burnett, A. 2012. Reproducible research. *CORE Issues in Professional and Research Ethics* 1(6).

Yale Law School Roundtable on Data and Code Sharing. 2010. Reproducible research. *Computing in Science & Engineering* 12(5):8–13.