Ethical Considerations of Sharing Data for Cybersecurity Research

Darren Shou

Symantec Research Labs, Culver City CA 92030, USA,
darren_shou@symantec.com

Abstract. Governments, companies, and scientists performing cyber security research need reference data sets, based on real systems and users, to test the validity and efficacy of the predictions of a given theory. However, various ethical and practical concerns complicate when and how proprietary operational data should be shared. In this paper, we discuss hypothetical and actual examples to illustrate the reasons for increasing the availability of data for legitimate research purposes. We also discuss the reasons, such as privacy and competition, to limit data sharing. We discuss the capabilities and limitations of several existing models of data sharing. We present an infrastructure specifically designed for making proprietary operational data available for cyber security research and experimentation. We conclude by discussing the ways in which a new infrastructure, WINE, balances the values of openness, sound experimentation, and privacy by enabling data sharing with privacy controls.

Keywords: data sharing, ethics, security

1 Introduction

1.1 Data Needs of Cybersecurity Research

Real world data is necessary for research and there is broad consensus in the security research community on what kind of data is needed [4]. Access to the large scale datasets needed for security research is limited primarily to the organizations that curate information for operational use in products. This access limitation is primarily due to intellectual property and privacy risks. These practical concerns are pitted against the ethical principle that access to data should be open and the scientific need for data to confirm experimental predictions. An increased availability of data would increase the research that could be performed and the usefulness of the results. Done properly, the availability of common datasets would enable peer review of cyber-security research. For the above reasons, increased data sharing has the potential to improve cybersecurity research. In this paper we examine the practical and ethical aspects of data sharing, discuss the capabilities and limitations of several existing models of data sharing and propose a model of data sharing for data generators. An example infrastructure that accommodates the concerns of this model is titled the Worldwide Intelligence Network Environment (WINE).

1.2 Existing Data Sharing Models

Companies and operators currently share data with academic and institutional researchers in several ways. Three approaches for data sharing are: using interns as data envoys, ad hoc sponsored research and data clearinghouses. Companies often hire interns from academic labs to experiment on the company's data. By using interns, companies maintain intellectual property rights and diminish the risk of data leakage. Unfortunately, the scale of using interns is limited by funding and the interns' time. As a result, the duration of such a data sharing project is often too short to accomplish significant results. Companies often contract research with university groups in order to share data. Again this does not scale well and benefits a relatively small group of researchers. Data clearinghouses like the Internet Traffic Archive collect datasets and make them publicly available for researchers. We list several specific examples below and elucidate their capabilities and limitations.

The Internet Systems Consortium provides a private information sharing framework, the Security Information Exchange (ISC SIE). It allows participants to contribute live feeds to be consumed by other members [10]. A limitation of this model is that does not provide a data preservation mechanism.

The Department of Homeland Security maintains a data archive known as PREDICT, the Protected Repository for the Defense of Infrastructure Against Cyber Threats. PREDICT acts as a clearinghouse between data providers and researchers [7]. A limitation of PREDICT is that data providers can retire datasets, making it impossible to reproduce past experiments.

The Internet Measurement Data Catalogue (DatCat) model is a searchable registry of donated data [9]. The DatCat model promotes reproducible research since researchers can cite the dataset handle in their research. Unfortunately, the database is down indefinitely.

The Internet Traffic Archive (ITA) contains mostly filtered network traffic traces [11].

The SIE, PREDICT, DatCat, and ITA models for data sharing have certain valuable capabilities but are limited in that none of them are data generators; they all rely on others for data contribution. Relying on donations from data generators is fundamentally problematic since data generators such as network operators and software companies often view the data they produce as intellectual property and a competitive advantage. Yet, a large amount of the interesting data needed by security researchers is collected, curated and preserved by such companies. Thus, a model for data generators to safely share their operational data without giving up control to a data collector would benefit the research community.

2 Ethical Considerations of Data Sharing

A number of statutes govern the legality of certain activities related to conducting cyber-security research [1][2][3]. However, ethical actions are not the same as legal actions; there exist legal activities that are not ethically permissible. The ethics addressed in this paper refer to well-founded standards of right and wrong. In the rapidly evolving field of data gathering it is possible to amass information on billions of people around the world in seconds. Not only must researchers ask whether their experiment is scientifically valuable, they must consider certain ethical questions. Researchers should ask, "Is this the right way to go about doing this experiment?" One can frame the applied ethics relating to norms in research, such as knowledge, truth and the avoidance of error. In this paper we will examine three specific aspects of scientific ethical concerns: openness, privacy, and sound experimentation.

2.1 Openness

By openness we mean the ethical principle that data, results and ideas should be shared and made available for peer review. However, access to datasets needed for openness in cyber-security research is limited primarily to the organizations that curate information for products and services.

Consider a computer security researcher that has proprietary access to a massive network trace dataset and conducts research that identifies a particular threat and a novel approach to addressing the threat. Should the author share the dataset although doing so may allow others to conduct research she might have interest in: and allow them priority and recognition [15]? The central question is how to weigh the benefit of the additional research that will follow from her sharing of the data against her personal ambitions. Her sharing the dataset with other researchers would serve not only the advancement of knowledge but also the public security interest.

Practical financial considerations must be addressed when companies practice scientific openness. First, can private data be liberated in a way that it is truly democratically available? Should a public company with petabytes of operational data be compelled to provide this data to its competitors as well as to educators and researchers? If so, the costs of providing large amounts of data are not inconsequential; who should bear the costs? A potential compromise is for public companies to make data broadly available for non-competitive purposes such as educational research. If an academic team has an idea for improving detection rates of say, malware, the team may use company data in their experiments. In this case, the academics would own their inventions but not the shared data. The aforementioned process addresses competitive issues and also suggests a way to deal with the costs. If a certain dataset is useful if it is broadly available, then the government could support the incremental costs of data sharing (i.e. hardware). This would be analogous to necessary software and travel expenses currently borne as expenses for scholarly pursuits. Government will in turn reap the technology benefits that the availability of real-world data will invigorate.

2.2 Balancing Privacy Rights

There is another ethical value in seeming conflict with the principle of openness: privacy. While openness calls for sharing data, tools, ideas and results, the principle of openness must be carefully balanced with a need for privacy. Indeed, much of the data needed for critical cyber-security research relies on data from real networks and users. For example, intrusion detection is dependent on large volumes of traffic so that researchers may generate signatures that minimize false positives and false negatives. There are of course several privacy laws that limit access to network traffic or address the storage of this information. In the US, there is the Wiretap Act that prohibits interception of the contents of communications, the Pen/Trap statute that prohibits real time interception of the non-content, and the Stored Communications Act that prohibits providers from knowingly disclosing their customer's communications [1][2][3]. In contrast to HIPAA, which restricts disclosures of health information but provides means for researchers to obtain information with and without individual consent, these cyber privacy laws contain no research exceptions. Ethically, what research actions such as those involving traffic monitoring and analysis ought be allowed?

Data curators should strive to be responsible stewards of the information they hold, protecting the ability for people to seclude themselves or reveal their information selectively. The use of the data should also be restricted to activities that provide value back to those whose information was volunteered. As we argued in the last section, open access to data is essential for research. But a balance needs to be struck between the scientific interest and the need to protect individual's privacy. There are good reasons for maintaining secrecy in research, from intellectual property protection and credit, to protecting research participants. Likewise, similar good reasons for openness have been discussed such as confirmation, progress, and transparency. The conflict in security research exists mainly because sensitive personally identifiable information may be present and there are those that would expose or use this PII in ways that do not benefit those that volunteered data. Technology can help with this balancing act, namely data handling tools such as anonymization and data leak protection (DLP) tools. Carefully anonymized datasets are useful since they reveal very little about individuals while still allowing researchers to learn from the data. However, the possibility of re-identification is changing the belief that perfect anonymization is possible [6][13]. The risk can be mitigated by coupling data handling techniques with the limited public release of information to trusted parties [13]. That is, those wishing to share private data publicly do so by restricting the data access to be on-site where data handling can be strictly enforced and motive may be further examined. Rarely do academic researchers have the motive to re-identify people in data as part of their experiments and so data curators may seek to prefer to share data with trusted researchers. In sum, data curators should obtain informed consent, evaluate benefit, use data handling best practices, and limit sharing to trusted relationships.

A specific schism between openness and private data is the desire for academics to publish and the companys desire for secrecy of data that may be viewed as intellectual property. Responsible disclosure practices can provide guidance for how to report insights that might incite as well as restrictions on revealing personally identifiable information[17], [8]. And companies ought to embrace open innovation practices such as data sharing to unlock previously undiscovered customer value from their operational data[5]. In sum, it is possible for companies and operators to strike a balance between privacy and openness. Moreover, the benefit of increased research to science and the public relies on the research being done according to proper scientific principles beginning with sound experimentation.

2.3 Enabling Sound Scientific Experimentation

In the first section we described how the availability of data is necessary for additional scientific experimentation. Experimental results must be independently confirmable if they are to be accepted by the scientific community and useful to commercial enterprises. Peer review and reproducibility are fundamental elements of the scientific method; these are the primary methods for identifying flaws in scientific research; everything from falsified data to statistically insignificance or misleading results. Scientific peer review is a self-correcting mechanism that eventually catches those that try to cheat the system, but it is imperfect; misleading, erroneous or fraudulent research can go undetected for years.

Confirmation is the best guard against flawed science and fraud. Two particular causes of flawed scientific research are the use of inadequate data sets and experimentation on data that is not archived for future access. If researchers do not have access to the appropriate data, then they cannot criticize fully or make comparisons between competing claims. Furthermore, if a given technology is only tested on a dataset that is knowingly orders of magnitude smaller than what is possible, then is any resulting error misconduct or accidental? Accidental experimental dataset errors will be reduced if scientists have access to the most comprehensive datasets as reference sets. The availability of such data sets would allow researchers to make fruitful comparisons between competing mechanisms, broadly measure progress, and validate or refute the claims of others. An example of this is the National Science Foundation policy that researchers must archive their data and methods so that others may test the methods and data [12]. In sum, the availability of archival data is essential for experiments to be verified through reproduction and for reliability to be measured with statistical analysis.

Practical limitations to sharing data include the dataset sizes and costs of the infrastructure. With peta-scale datasets necessary for research, the datasets are not easily replicated. Making data universally accessible would be ideal but is not viable considering the computing and storage demands. One solution is to use a review board to limit access to data resources when costs are not underwritten. A review board would also aid quality control by ascertaining the standing and originality of research plans.

ITA and similar aforementioned models address the cost issue from a technical standpoint well since they offer to maintain a central repository of data for multiple researchers. However, given the risks associated with intellectual property and proprietary information facing operators and companies, it is more likely that most companies will want to host their own datasets. Furthermore, if an operator restricts to onsite access only, it can provide more than just data. It can provide computing resources, subject experts and experimentation facilities. Having researchers onsite with companies' datasets encourages cross-fertilization of ideas amongst researchers and employees, potentially resulting in increased commercial technology. Moderated control of data sharing is a reasonable method for companies to deal with the expense of sharing and also the competitive and privacy issues previously discussed. Unfortunately, for companies to limit access to their data this way conflicts with the principle of openness and scientific peer review.

3 Conclusion

In this paper we have discussed the importance of a model for data sharing that provides scientists reference data for confirmation while protecting the privacy of those represented in the data set. Symantec's Worldwide Intelligence Network Environment (WINE) is an existing implementation of such a data sharing model. WINE addresses two related shortcoming of the various existing data sharing models, SIE, PREDICT, DatCat, and ITA; these models rely on volunteered data and the continued availability of the data is subject to the whims of those that volunteered the data.

WINE provides academics with access to precisely those security related data feeds that many data generators choose not to volunteer. WINE makes available Symantec telemetry data from over 75 million participating machines, including every attack Symantec finds on both the file system or network side as well as suspicious files or traffic that are likely threats. Such attack data includes a rich set of metadata including anonymized attacking addresses, OS version, process name, geographic local, language, URL the file or attack came from, etc. In addition, 5.5 million malware, 100,000 spam emails, and 60 TB of binaries' metadata encountered over years of anonymous submissions is gathered from millions of sensors, honeynets and decoy accounts: [16]. Where applicable, tools, scripts, and documentation will also be archived with datasets. And there are plans for visualization and analytical tools. Furthermore, WINE retains datasets indefinitely, as permitted by cost and legal restrictions. This allows scientists to reproduce past experiments and compare the effectiveness of older algorithms to newer ones.

In the WINE model, researchers browse a catalogue of datasets and construct a proposal along with a data request. The validity of the proposals and the availability of the requested data are evaluated by an advisory board of external and internal researchers. The intellectual property developed by the researchers using WINE is theirs and they are encouraged to publish their results responsibly. We sincerely hope that, for the benefit of cyber-security research, other companies choose to establish models similar to the guidelines set out in this paper.

References

1. 18 U.S.C. §2510-2522.
2. 18 U.S.C. §2701-2711.
3. 18 U.S.C. §3121-3127.
4. Camp, J., Cranor, L., Feamster, N., Feigenbaum, J., Forrest, S., Kotz, D., Lee, W., Lincoln, P., Paxson, V., Reiter, M., Rivest, R., Sanders, W., Savage, S., Smith, S., Spafford, E., and Stolfo, S. : Data for Cybersecurity Research: Process and Wish List. National Science Foundation Workshop on Cyber Security Data for Experimentation. (2010)
5. Chesbrough, H.: Open Business Models: How to Thrive in the New Innovation Landscape. Harvard Business School Press. Boston. 2006.
6. Coull, S., Wright, C., Keromytis, A., Monrose, F., Reiter, M.: Taming the Devil: Techniques for Evaluating Anonymized Network Data. Proceedings of the Network and Distributed System Security Symposium, NDSS 2008, San Diego, California
7. Department of Homeland Security: Protected Repository for the Defense of Infrastructure Against Cyber Threats https://www.predict.org
8. Google: Security. http://www.google.com/corporate/security.html
9. Internet Measurement: The Internet Measurement Data Catalogue (DatCat) http://imdc.datcat.org
10. Internet Systems Consortium: Security Information Exchange. https://sie.isc.org
11. Internet Traffic Archive: The Internet Traffic Archive (ITA) ita.ee.lbl.gov
12. National Science Foundation: Dissemination and Sharing of Research Results: NSF Data Sharing Policy http://www.nsf.gov/bfa/dias/policy/dmp.jsp
13. Ohm, P.: Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. SSRN eLibrary, 2009.
14. Powner, D.A. , Wilshusen, G.C: Key Challenges Need to Be Addressed to Improve Research and Development. Technical Report, GAO-10-466 (2010)
15. Resnik, D.B.: What is Ethics in Research and Why is it Important? National Institute of Environmental Health Sciences. http://www.niehs.nih.gov/research/resources/bioethics/whatis.cfm
16. Symantec Corporation: Internet Security Threat Report. Technical report, Symantec Managed Security Services. (2010)
17. Symantec Corporation: Symantec Responsible Disclosure Policy. http://www.symantec.com/research/Symantec-Responsible-Disclosure.pdf