Ethics in data sharing: developing a model for best practice

Sven Dietrich*, Jeroen van der Ham†, Aiko Pras‡, Roland van Rijswijk-Deij§, Darren Shou¶, Anna Sperotto‡, Aimee van Wynsberghe‡, Lenore D. Zuck**
* Computer Science Department
Stevens Institute of Technology, Hoboken, NJ, USA
spock@cs.stevens.edu
†Faculty of Science, Informatics Institute
University of Amsterdam, Amsterdam, The Netherlands
vdham@uva.nl
‡Centre for Telematics and Information Technology (CTIT)
University of Twente, Enschede, The Netherlands
a.pras,a.sperotto,A.L.vanWynsberghe@utwente.nl
§SURFnet bv
Utrecht, The Netherlands
rijswijk@surfnet.nl
¶Symantec Research Labs
Culver City, CA, USA
darren_shou@symantec.com
**Computer Science Department
University of Illinois at Chicago, Chicago, IL, USA
lenore@cs.uic.edu

***Abstract*—As an outcome of a seminar on the 'Ethics in Data Sharing', we sketch a model of best practice for sharing data in research. We illustrate this model with two current and timely real-life cases from the context of computer and network security.**

***Index Terms*—ethics; data sharing; best practice; computer security.**

I. Introduction

In January 2014, the renowned Dagstuhl Seminars in Computer Science brought together computer scientists, an ethicist and legal scholars to discuss the topic of "ethics in data sharing" [1], [2]. Three main themes requiring ethical attention were identified by this group of researchers:

- Best Practices and Institutional Review Boards (IRBs) for Ethics in Computer Science,
- Models of Ethics in Producer-Consumer Relations in Data Sharing for Research and Operations, and
- Building Ethical Technology.

The discussions on the first two themes eventually converged, and it is the topic of this report.

II. Goals

The goal of the group was to sketch out a model for best practice when it comes to data sharing and maintaining an ethical standard for doing so. The researchers began by mapping out the stages of the design process during which data is collected which can eventually be shared. The group took as their starting point the computer ethics concept of embedded values [3]. This concept asserts that there are values embedded in a technology such that when a technology is used it will promote said value.

With this in mind, the design process is an integral part for sketching the values to be embedded in a technology. Accordingly, the group set out to investigate the relationship between values along the stages of research involving data collection for sharing. The discussion on what is data, research (whether academic or not) on collected data has at least the following stages:

1) define the purpose of the research;
2) design and implement the tools and experiments for the data collection and analysis;
3) collect the raw data (possibly by acquisition from a third party);
4) store the data;
5) analyze the data;
6) disseminate the results; and
7) curate the data.

Alongside the stages of data collection, of equal importance are the actors involved in said collection. There are multiple stakeholders in the process: Researchers, data collectors, research participants, organizations (from private companies to academic institutions), third parties, each with their own interests and ethical motivation. Each of those parties also assigns value to the research, be it actual profit, quest for truth, security, reputation, education, awareness, and many others.

Even with the best of motives and best attempts to act ethically and to avoid harm, there may be unintended consequences for data-centric research. We strongly recommend at each stage of the research to perform an (re-)evaluation of values, for each of the stakeholders, i.e., the researcher, the data provider, the subjects and any other third party that may be involved. The value analysis involves a discussion of the values intended by the researcher, the values of the users or society that might be impacted (positively or negatively) and the values that have been neglected or traded-off. Such a discussion requires a critical stance on the values made explicit as well as insight into how said values can be interpreted from the ethics perspective. For these reasons we advocate in favor of having an ethicist guide the analysis. For a more detailed description of what such a value analysis looks like see van Wynsberghe and Robbins 2013 and their discussion of the tasks of the ethicist [4].

This also entails an evaluation of the rights data subjects have, how they can exercise these rights, how they are debriefed in case of (possibly unavoidable) deceit, what are the potential harms, and how (and if) these harms can be mitigated. The evaluation should be performed by a third party, such as an IRB, an ethics adviser or others. It is conceivable that one would have to revise research plans, or even, in extreme cases, halt it at its midst (as was the case in Stanford's famous Prison Experiment.)

It is also important to understand the assumptions of transitivity of rights and responsibilities along the path of the shared data. That is, what did the originator agree to for sharing the data, and what could violate those assumptions later? How are these properly passed down while protecting the rights of the source?

Throughout this short report we assume that at least researchers and data providers are motivated to behave ethically. This is a naïve assumption (as any spammer, let alone those who recently released potentially harmful documents, will attest to), yet, a global system of incentives for ethical behavior for the world at large, or even for researchers who struggle to "publish it first," is well beyond the scope of this report.

To summarize the above, we suggest that the researchers, guided by an ethicist, will engage in value analysis at each stage in the research and design process. The value analysis will criticise the intended values and will look for value trade-offs. The value analysis can be from the perspective of the researcher, a third party or a future user. The perspective that one takes will change the interpretation and prioritization of values. We looked at several case studies and analyzed them according to the outline above. Below is the analysis of two of them that represent somewhat opposing ends: The effectiveness of attempts to block The Pirate Bay (TPB) in the Netherlands [5] conducted by one of the authors of this document, and the infamous Internet Census 2012 [6] that used a self-developed botnet to scan the entire Internet.

Table I provides a brief ethical analysis using the proposed analysis framework. It is beyond the scope of this paper to present an in-depth ethical analysis or a detailed justification for the framework but this will be the goal of future work. The table here only includes the point of view of the researcher, for a complete analysis also the point of view of other stakeholders should be included. To be clear, the table presents the values that were raised for discussion but it does not go into a detailed account of how they are defined, interpreted, or prioritized. This will be the subject of on-going and future work.

Notable differences are that the Internet Census 2012 data is hard to verify, has minimal accountability (the author decided to hide behind a PGP public key), and defies privacy. In contrast, the goal of the study on blocking attempts for The Pirate Bay was to study effectiveness of such blockades. In so doing it aimed at collecting accurate, reliable data but risked minimizing the privacy of users. The blocking study provides considerable (though incomplete) data provenance, and the data it gathered is not available to the world at large.

The Internet Census 2012 data is an interesting example of research with potential to raise awareness of vulnerable devices and malware. Yet, it reveals much privacy invading information, and can not be verified. In contrast, the Symantec Worldwide Intelligence Network Environment (WINE) model [8] provides researchers with means to obtain malware data, conduct reproducible experiments, provides for data provenance, yet avoids the pitfalls of publishing privacy invading information.

While performing the ethical analysis we have also been forced to conclude that it is almost impossible to do this properly without the input of the researcher. Many values are not expressed explicitly, and can only be extracted by explicitly performing the analysis as proposed. For the Internet Census 2012 the researcher was not present and we have resorted to extracting these values from the description of the research and its motivation.

The ethical analysis presented here has shown to be a good toolkit to use in a dialogue between the researcher and an ethicist. The framework helps to structure the discussion, to cover all the facets of performing an experiment where data is collected. This can be done prior to presenting a proposal to an IRB, and will hopefully also help the IRB to quickly assess the morality of the research proposal. Other examples of building up ethical awareness can be found in [7].

III. Follow-up efforts

One of the drivers for the discussion on the ethics in producer-consumer relationships was and is a shared desire by a number of participants to the Dagstuhl seminar to share more data – specifically for network and network security research – and to share it more openly. Two examples of current practices for sharing this kind of data were presented at the seminar, these are:

- The Symantec WINE model [8], mentioned above; briefly summarised, this model works such that researchers can apply for access to the Symantec repository of data on network security threats and incidents that it collects in the course of its day-to-day operations. Researchers that are accepted to the program are invited to visit a data haven on Symantec premises and can execute their research algorithms on hardware provided by Symantec. The goal of the program is to be inclusive and open.
- SURFnet, the National Research and Education Network in The Netherlands, shares data with academic researchers on a regular basis. This data is often aggregate data (network flow information) but can also consist of full network traffic captures for certain protocols and services. There are clear ethical concerns for this type of data sharing, since *a)* the data being shared may contain personally identifying information about users of the SURFnet network, and *b)* it is hard or impossible to allow users to opt-out of this data sharing. Current practice to address these concerns is to only share data with "trusted" researchers (often this means having a personal relationship with the researchers) and to scope the data

TABLE I
A MODEL FOR ETHICS IN DATA SHARING

	TPB Blockade Effectiveness	Internet Census 2012
Concept and Design	Design and implementation of the tools and experiments *Values: accountability, objectivity, fairness*	Port scanning with the use of middle nodes, changed over time to minimize bandwidth usage/ load, did not change passwords, did not erase disks, removed after reboot - minimized impact *Values: Non-maleficence, transparency, fairness, security, privacy, truth*
Data collection	Running the measurements, participating in the data exchanging process *Values: Truth, safety, objectivity, beneficence, transparency of tool, however not for the user*	Collection of data without harming the target system, creating bots, installed software, invasion of open systems *Values: as above*
Data storage	On an encrypted local disk *Values: Privacy, reputation, truth, accountability*	Most efficient way (technology perspective) *Values: Efficiency and effectiveness*
Data Analysis	Geo Location full data; IP to AS mapping through a third party service, aggregation and statistical analysis *Values: Objectivity, truth, accountability*	Hilbert curves, geographical distribution, standard analysis *Values: Objectivity*
Data Verifiability	Manual verification with random sampling *Values: Weighing of effectiveness and efficiency against full data analysis*	None
Dissemination	Publications, outcome in a technical report (public after review by lawyers) *Values: Truth, accountability*	Data on Web site, interpretation/results and full data set online *Values: Secrecy, awareness of security*
Data Curation	Stored offline; shared only aggregated data. *Values: accountability, privacy, truth*	Data shared publicly without warning *Values: None*

sharing under a Non-Disclosure Agreement (NDA) that clearly defines:

- What data is shared
- For what purpose the data may be used
- Who has access to the data
- How long the data may be stored and when it must be destroyed
- Conditions of publication (e.g. references to individual IP addresses must be anonymised)

Both examples show that the organisations involved take ethical concerns seriously. Nevertheless, both approaches have limitations given the desire to share data more broadly. In the first case, requiring researchers to come to a data haven may be prohibitive both because international researchers will incur a greater cost for having to travel to the data haven, and also because having to execute research algorithms on systems provided by the owner of the data haven means researchers will have to hand over what may be their core intellectual property to a third party (note that this can also be seen as a benefit, since it can balance the trust relationship between the data provider and data consumer). In the second example, the problem lies in the fact that current data sharing practices rely on personal relationships. This is problematic, for instance with a view towards reproducibility. If, say, data is shared with a friendly research group within the SURFnet constituency and another, unaffiliated and unknown research group from another country wants to reproduce research results based on the same or similar data then there is currently no good way to establish the required trust and to assess the risks and ethical implications.

Recall that there was a desire by participants to the seminar to share more data and share it more openly. Clearly, the examples above go a way towards realizing that goal but there is definite room for improvement. To address this, a number of participants to the Dagstuhl seminar have taken the lead to work towards a policy framework for data

sharing that is intended to help producers (i.e. parties that have data and are willing to share this for research purposes) and consumers (i.e. researchers) formalize their relationship and to address the ethics as well as the legal aspects of sharing data. Questions we intend to address are:

- Are the NDAs such as those used by SURFnet sufficient or do they require extra clauses?
- How best to review the ethical considerations from both the producer as well as the consumer side of the data sharing relationship such that conflicting situations become apparent and can be addressed adequately.
- Should there be a review board on both sides of the relationship and if so, how will these interact?
- How best to give a voice to users (who may be subjects of the research unbeknownst to them) affected by the research while not diminishing the quality of the data for research purposes?
- How to guarantee maximum transparency and accountability?
- How to establish a thorough, responsible process without getting bogged down in endless procedures?

The first session is scheduled to take place later in March 2014. We intend to publish the first outcomes of our ongoing discussions later this year.

ACKNOWLEDGMENTS

The authors would like to thank the participants of the Dagstuhl seminar 14052 for their valuable input, and the organizers at the Schloss Dagstuhl Leibniz Center for Informatics. This work has been partially supported by the EU FP7 Flamingo Network of Excellence (ICT-318488).

REFERENCES

[1] Dagstuhl Seminar 'Ethics in Data Sharing'. http://www.dagstuhl.de/en/program/calendar/semhp/?semnr=14052.

[2] Roland van Rijswijk-Deij (2014). Ethics in Data Sharing Seminar Blog. https://blog.surfnet.nl/?p=3174, January 2014.

[3] Helen Nissenbaum (2001). How computer systems embody values. Computer, 34(3), pp 120–119. DOI:10.1109/2.910905.

[4] Aimee van Wynsberghe, Scott Robbins (2013). Ethicist as Designer: a pragmatic approach to ethics in the lab. Science and Engineering Ethics. DOI:10.1007/s11948-013-9498-4.

[5] Joost Poort, Joma Leenheer, Jeroen van der Ham, Cosmin Dumitru (2014). Baywatch: Two Approaches to Measure the Effects of Blocking Access to the Pirate Bay, Journal of Telecommunications Policy, January 2014. DOI:10.1016/j.telpol.2013.12.008.

[6] Internet Census 2012, Port scanning /0 using insecure embedded devices, Carna botnet, http://internetcensus2012.bitbucket.org/paper.html.

[7] David Dittrich, Michael Bailey, Sven Dietrich (2011). Building An Active Computer Security Ethics Community, IEEE Security and Privacy Magazine, July/August 2011. DOI:10.1109/MSP.2010.199.

[8] Darren Shou (2011). Ethical considerations of sharing data for cybersecurity research, Proceedings of the 2011 Workshop in Ethics in Computer Security Research, in Financial Cryptography and Data Security, pp 169–177, Springer LNCS 7126 (Editors: George Danezis, Sven Dietrich, Kazue Sako). DOI:10.1007/978-3-642-29889-9_15.