Ethics and Internet Measurements

Jeroen van der Ham[1,2] and Roland van Rijswijk-Deij[3,4]

[1] *Technical University Delft, The Netherlands*
[2] *National Cyber Security Centre, The Netherlands*
[3] *University of Twente, The Netherlands*
[4] *SURFnet bv, The Netherlands*
Email: jeroen.vanderham@ncsc.nl; r.m.vanrijswijk@utwente.nl

Received 15 September 2017; Accepted 18 September 2017;
Publication 6 October 2017

Abstract

Over the past decade the Internet has changed from a helpful tool to an important part of our daily lives for most of the world's population. Where in the past the Internet mostly served to look up and exchange information, it is now used to stay in touch with friends, perform financial transactions or exchange other kinds of sensitive information. This development impacts researchers performing Internet measurements, as the data traffic they collect is now much more likely to have some impact on users.

Traditional institutions such as Institutional Review Boards (IRBs) or Ethics Committees are not always equipped to perform a thorough review or gauge the impact of Internet measurement studies. This paper examines the impact of this development for Internet measurements and analyses previous cases where Internet measurements have touched upon ethical issues. The paper proposes an early framework to help researchers identify stakeholders and how a network study may impact them. In addition to this, the paper provides advice on creating measurement practices that incorporate ethics by design, and also considers the role of third-party data suppliers in ethical measurement practices.

Keywords: Ethics, internet measurement, ethical review, ethics committee, privacy, IRB.

Journal of Cyber Security, Vol. 5_4, 287–308.
doi: 10.13052/jcsm2245-1439.543

1 Introduction

Our daily activities increasingly involve the Internet in some way. This means that more and more data is sent over the Internet, and also that more and more *sensitive* information is sent over the Internet.

This impacts Internet measurements for cybersecurity as well, as in many cases it is hard to distinguish beforehand between sensitive and non-sensitive information. It is even possible that the results of measurement studies themselves become sensitive information.

The topic of the ethics surrounding network measurement research has surfaced several times in the history of Internet measurement. Yet it has not received much attention so far, and deliberations on sensitivity of information, or moral aspects of experiments are still too often not part of the final publication.

In 2015 debates over general measurements on the Internet garnered a wider interest. There was outrage over the way Facebook performed an experiment on filtering of timeline messages [1]. And closer to the measurement community, there was a debate over experiments on measurements of censorship at the 2015 ACM SIGCOMM conference [2, 3].

In this paper we investigate ethics in the context of network measurement research from several angles. We start out in Section 2, by examining cases of past studies where ethics played a role and where, in hindsight, a thorough ethics review before the research was performed may have been warranted. Next, in Section 3, we consider past efforts and related work in incorporating ethics into network measurement research. Then, in Section 4 we provide a framework for ethical analysis. This framework not only specifies guidelines for IRBs or Ethics Committees, but also includes advice on designing ethical measurements from the ground up and touches on the role of third party data suppliers. Finally, in Section 5 we summarise our conclusions and outline areas that require future study.

2 Case Descriptions

This section discusses a selection of cases from recent years, in which ethics have played a role. The cases are relevant in different ways to the Internet Measurement community. We have structured each case study such that it first describes the research itself, followed by a summary of the ethical discussion surrounding the research and ending with an analysis of the causes for the discussion.

2.1 Facebook Filtering

In 2014 a paper was published entitled 'Experimental evidence of massive-scale emotional contagion through social networks' [1]. This paper describes how filtering of messages with either positive or negative emotional content can impact the writing behaviour of users of a social network. To support this conclusion an experiment was conducted in early 2012 on more than 600,000 Facebook users.

The authors of the paper did not seek permission from their Institutional Review Board. As the university researchers did not come into contact with the data itself but only provided the analyzation methodology, they saw no need to do this. Furthermore, the analysed data was a pre-existing dataset, i.e. the data had already been collected before the university researchers became involved in the research.

2.1.1 Discussion

The publication was discussed on Twitter several weeks after publication at the end of June 2014. Fairly quickly commenters discovered that Facebook had performed an experiment on unwitting users, which attempted to influence their emotional state. Many participants of the discussion were outraged by how the experiment was performed and their lack of transparency on it[1]. Just several days later Cornell University issued a press release[2] explaining their involvement.

The discussion around this study prompted other researchers to examine the challenges around Internet-based research. One article focuses on these challenges and examines policy in the UK [4]. Their conclusion was that in 2014 at least in the UK there were no formal guidelines for this type of research.

2.1.2 Analysis

A study on emotions, especially when this involves a popular service such as Facebook, has a high probability of getting attention. The experiment performed on Facebook users by Facebook employees is in itself legal[3].

[1] URL: http://www.theatlantic.com/technology/archive/2014/06/everything-we-know-about-facebooks-secret-mood-manipulation-experiment/373648/

[2] URL: http://mediarelations.cornell.edu/2014/06/30/media-statement-on-cornell-universitys-role-in-facebook-emotional-contagion-research/

[3] There is, however discussion on the legality of using the data for scientific research since the Facebook User Agreement did not contain the 'research' purpose until 2014.

However, academic research uses a higher standard for performing research. This includes reviews of human subject research by IRBs or ethics committees, and the principle of informed consent.

The IRB was not involved in this case as the data had been collected outside of the university and no university employees came into contact with the data. According to the guidelines, an IRB review was not necessary for pre-existing data. This has been described by some as 'IRB laundering' of experimentation data [5].

The subjects in the experiment were selected by Facebook and were not approached beforehand to ask for consent. In academic practice it may be permissible to seek consent *post facto*, for instance when knowledge of an experiment could influence the result. The participants in this case were, however, not even informed afterwards about their participation, let alone about the results of the study.

2.2 Censorship Measurements

In 2015 the program committee of the ACM SIGCOMM conference had many questions about a paper that discussed censorship measurements [2]. So much so that the published paper includes an official statement from the program committee, calling for ethical guidelines on measurement papers.

The paper describes an experiment where certain websites are extended with measurement scripts. These measurement scripts make the browser perform certain requests for websites. This is done in such a way that an observer can also record whether this request has been performed successfully or not.

2.2.1 Discussion

The program committee reviewing the paper raised ethical concerns with the paper. The submission guidelines of SIGCOMM requires authors to engage with their IRBs. The first author of the paper contacted their IRB to review the experimental approach. This IRB concluded that the experiment did not involve human subjects, and consequently that there was no reason for a formal IRB review. The second author of the paper, meanwhile, had moved to a different university, so they requested a review by the IRB of that institution. Again this IRB declined a formal review because they reasoned the study was not human subjects research.

The paper sparked a discussion in the program committee, the measurement community and even beyond that. This case has also been extensively reviewed by Narayan and Zevenbergen [6].

2.2.2 Analysis

The Institutional Review Boards of both universities appear to be ill-equipped to review this research. According to the statements, this is because of regulations surrounding the IRBs and the formal definition of human subject research [7, 8]. In many institutions, the IRB reviews proposed research and experiments to protect human subjects. This means an ethical review of an experiment is only done when human subjects are involved. In this case both IRBs ruled that the experiment did not involve human subjects.

The program committee on the other hand felt that this research did touch human research subjects, or at least could possibly affect people. This shows an important disagreement between the program committee and the IRBs, which raises many questions. First, there does not seem to be a way for a program committee to express concerns towards an IRB. Another question would be whether the current IRBs are technically capable of accurately identifying the human subjects component of this kind of research. Another possibility is that the researchers themselves were not able to either identify the human subjects component, or were not able to explain this completely to the IRBs.

The ethical issues in this paper, again, revolve around informed consent, which was not requested from (unknowing) participants in the experiment. The researchers argued that this was not done because it was difficult to explain, nor would it lower the risk to the user. The Program Committee was also worried about possible actions against the users browsing the websites extended with the measurement scripts. It is not unthinkable that requesting many censored websites will garner attention from law enforcement in regimes with censorship.

2.3 NYC Taxi Dataset

New York City requires that taxis keep detailed logs of their rides, including the pick-up and delivery GPS-location, as well as the price and tip received for each ride. This data is recorded by the city council and is available upon request[4].

A citizen researcher requested the data and received an anonymized version of the data on a USB drive. However, it turned out that the data was anonymized

[4] Since the data is collected by the City of New York in the US, it is subject to Freedom of Information Law (FOIL) requests. FOIL is the local variant of the (federal) Freedom of Information Act.

using simple MD5 hashing. In addition, the source data, the taxi vehicle and hack license numbers did not contain enough entropy to make this a secure way of anonymization. This enabled the researcher to re-identify all of the taxi data.

To make matters worse, once the taxi numbers in the dataset were re-identified, it was also possible to correlate the information with already existing data. Celebrities living or visiting NYC are often photographed, also when they enter or exit taxis. Published photos often including timestamps, which makes it possible to correlate these photos back to a specific record in the dataset.

2.3.1 Discussion

While the case described above does not directly deal with Internet traffic measurements, it clearly shows that data is not always neutral, even if it is anonymised. Anonymisation is not easy to do, especially with low-entropy source data. This is even more so when this kind of data can be correlated to other existing datasets. The taxi dataset is an obvious example of how this can go wrong.

Internet measurement data can exhibit the same qualities. For example, IPv4 addresses have very little entropy, and with a reasonably accurate timestamp this data can easily be correlated with other existing or observed datasets [9, 10]. As is the case for the taxi data, it is very hard to predict how this data can be correlated or how this would impact subjects.

2.3.2 Analysis

This case shows another area that may not currently be on the radar of IRBs. An attempt was made to anonymise the data, but it is apparent that insufficient expertise was available to do this securely. The data in the taxi set is very similar to data that is available in the measurement community: the source data has low entropy, and many related sources are available to enrich and deanonymise the original dataset.

A similar case happened in the social sciences in 2008 with the release of the ‘Taste, Ties and Times’ dataset [11]. A group of researchers published an anonymised dataset of a cohort of students at an American university. The data was very quickly deanonymised, exposing private information about students that was not generally available [12]. This study was actually reviewed by an IRB at the time, which did not see any issues in releasing this information.

2.4 Blockade Measurement

In 2012, a court in The Netherlands ordered several ISPs to block access to the PirateBay website, a search engine that mostly indexes pirated content. The effectiveness of this blockade was subject of debate in the court-case, there were specific concerns about the proportionality of the measure. This could not be directly measured by the ISPs themselves, as they are not allowed to inspect their user's traffic.

Unlike the ISPs, the University of Amsterdam did perform a measurement to establish the effectiveness of the blockade. Performing the measurement required collection of the IP addresses of downloaders, so that a distribution of users over the different Dutch ISPs could be established. Data was collected at moments before the blockade, during a partial blockade (two out of six major ISPs) and after an almost nation-wide blockade (all major ISPs). While a detailed discussion of the work is out of scope for this paper, we note that the study showed that the blockade was ineffective (i.e. did not stop users from downloading pirated content found through the PirateBay).

2.4.1 Discussion

The ethical aspects of this research are in recording the IP addresses for the purpose of this measurement. This normally requires permission from the owner, and things are made worse by the fact that the measurement possibly records an illegal act by the downloader. Concurrent with this measurement study, a survey was performed to measure the usage of BitTorrent by the population. Neither results are conclusive by themselves, however both gave very similar results, strengthening the conclusion that the blockade was ineffective [13]. A more extensive review of the ethical aspects of this case is described in [14].

At the time there was no IRB or ethics committee to review this experiment at the University of Amsterdam. There was no support at the university itself to guide on this issue. The author reached out to the ethics advisor of the University of Twente, who performed an ethical review of the study (*post facto*).

Before this case, there was already some movement to start an ethics committee for computer science at the University of Amsterdam. This case has helped shape this effort. The author has also helped start an additional ethics committee for one of the Master programs that often deals with digital security projects [15].

2.4.2 Analysis

The blockade measurement case has many different stakeholders: *the clients* downloading copyrighted materials, who are identified through this experiment; *the copyright industry* which is afraid for their income; *the ISPs* which would have to implement additional technical measures to block websites; *the researcher* performing the experiment, who must be able to demonstrate the results; *the general public* which may be harmed by increased censorship.

3 Related Work

The Menlo Report [16] was published following an initiative of the US Department of Homeland Security. The purpose of the initiative was to translate the principles of the Belmont Report (which deals with human subject studies in the medical and social sciences) to the ICT Research context, with possible additions. The companion report [17] to the Menlo report describes example cases and explain how ethical analyses of these cases are performed.

Partridge and Allman have made the case for an Ethical Considerations paragraph in Network Measurements publications [18]. They provide an overview of several developments in measurement publications, and call for ethical considerations paragraphs which are reviewed by program committees. The authors provide some generic advice promoting ethical awareness, but no directly applicable advice for authors or reviewers, other than referring to the previously mentioned Menlo Report.

'Forgive us our SYNs' [3] describes a measurement study similar to the study described in Section 2.2. At the same time the paper describes an extensive overview of the ethical considerations of the study itself, aiming to provide a more generic model.

Following a Dagstuhl Seminar in 2014, several researchers collaborated to propose a model for ethics in data sharing [19]. SURFnet, the National Research and Education Network in The Netherlands, followed up with this model and created a policy for data sharing, where ethical review plays an important role [20]. SURFnet's data sharing practice will be discussed in a brief case study included in Section 4.3.1.

4 Framework for Ethical Analysis

An earlier publication contains a description of embedding ethics in a systems and network engineering educational context [15]. The students are briefly instructed on ethics at the start of their curriculum. During the master programme, the students are required to write an ethical analysis for each of the projects that they perform.

For the masters program there is a (small) dedicated ethics committee that is in direct contact with the students. The ethics committee works with the students to analyse ethical aspects of their research, and also instructs them on how to improve their approach.

The model used in the education is an extension of the data sharing model as described in [19] and the Ethical Impact Assessment [21]. This means that the affected parties should be identified and the possible impact should be established. The affected parties are the users, the researcher, the university, any other related parties, as well as the general public. Many of the projects include security research, which may lead to discovery of vulnerabilities. Before the research is started, the students are forced to think about the impact of possible vulnerabilities for the vendor, the users, and the general public.

The proposed framework for Internet measurements follows the same approach, with a more extensive emphasis on the different stages of measurement research:

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. analyse the data;
6. disseminate the results; and
7. curate the data.

For each of the stages the relevant parties should be identified and care should be taken to accurately execute each of the steps. It is important to emphasise that the analysis for *all of the stages* is performed before the measurement is performed.

4.1 Ethics Review

The above framework identifies the important elements of traffic measurement research. With the framework researchers can identify relevant stakeholders and start gauging an ethical impact of the research on these stakeholders. The effort of a researcher should not stop there, however. Researchers should also actively involve an IRB or EC when a direct impact is found on stakeholders other than the researcher. This should be done regardless of whether the perceived effect is positive or negative.

The cases presented in this paper have shown evidence that computer science research, and especially traffic measurement research, has clear ethical impacts. This means that IRBs and ECs should be involved in reviewing this type of research. We therefore urge IRBs and ECs to re-evaluate their regulations and procedures to gauge whether they are actually able to execute such reviews if required, and whether additional training or education is required.

Researchers themselves must also be educated on the possible impact that their research may have. The cases in this paper show that research has changed in such ways that measurement experiments and data can have significant ethical and legal impacts. Many of the current researchers have had only limited education on ethics, and may not realise the impact their research may have.

At the same time, institutions should take a pro-active role in supporting researchers in their ethics review. Having an ethics advisor available to discuss an experimental design is extremely helpful. In many cases it can even help to design the experiment in such a way that it has no or minimal ethical impact on other stakeholders.

4.2 Ethical Measurements by Design

While an ethics review is a necessary step to take *before* traffic measurement research starts, it is equally important to incorporate ethical thinking into *how* traffic measurements for research purposes are performed. There are a number of best practices that have been published over the years. The oldest, most high-level, yet still relevant was documented in 1991 by Vint Cerf in RFC 1262 [22]. Most of the seven guidelines documented in that RFC are just as relevant today as they were in 1991. Next, Allman and Paxson address etiquette around reusing traffic measurement data shared by other researchers [10]. They address both problems in releasing datasets (how to describe data, conditions for use, . . .) and problems in re-using shared datasets (proper attribution, being mindful of the effort data sharing takes, . . .). Finally, Papadopoulos and Heidemann outline their views on best practices for active network measurements, arguing that the burden on performing responsible active measurements resides exclusively with the researcher, and that one may only hope, but cannot expect, that network operators are sympathetic to active measurement traffic [23].

While these best practices provide useful guidelines, they are disparate, and no single practical set of guidelines for ethical network traffic measurements exists. Therefore, in this subsection, we provide a *practical* set of guidelines for precisely that purpose[5]. This set of guidelines is by no means exhaustive; nevertheless, we believe it can support the design of ethically responsible network traffic measurements in many cases.

4.2.1 Passive vs. active measurements

We start by observing that the ethical considerations when designing network traffic measurements differ, depending on whether they will be passive or active measurements. While this may seem obvious to some, the decision process on what to consider in a measurement design should always start with this question. Generally speaking, for passive measurements we believe that the primary concern is the privacy of the users whose traffic is being measured, whereas for active measurements the primary concern is the impact of the measurement on the systems and networks that are measured. Below, we provide more detailed guidance for both types of measurement.

While the guidelines specified below assume that a clear distinction can usually be made, we note that it may sometimes be ambiguous whether a measurement should be considered passive or active. This must not be an excuse to disregard, e.g., risks to users privacy because a measurement is considered active rather than passive (where the underlying assumption is that if one actively has to query something to retrieve personally identifiable information, it should be considered public information). We make this distinction between passive and active purely to make the decision process on what guidelines to generally consider when designing measurements easier for researchers.

4.2.2 Guidelines for passive measurements

For passive measurements, we consider two things to be the most important. First of all *privacy*. Passive measurements typically rely on observing real-world network traffic, and may thus impact the privacy of the individual users whose traffic is being observed. In order to minimise the risk to user privacy by design, we recommend taking into account the following guidelines:

Discard privacy-sensitive data as early as possible – This starts with thinking about what information is needed to perform the research one is about to embark on. It may seem attractive to always store all collected data, in case

[5] The guidelines are based on a similar set of guidelines included in the Ph.D. thesis of one of the authors of this paper [24].

new insights arise as the research is ongoing. This is, however, a pitfall that can quickly lead to high risks to user privacy and scope creep with regards to the original reason data was collected. We therefore advocate the same approach that, for example, many forms of data privacy legislation require, which is to document *before* collection starts exactly *what* data is collected, for *which purpose* and over *which period*. If there are new insights that require additional data, this process should be revisited.

Only store aggregate or anonymised results – Measurement results are often stored long term, to facilitate reproducibility of research. Many institutions, nowadays, even mandate such long term storage and require researchers to submit data management plans. To protect the privacy of users in such cases, we strongly recommend storing only aggregate results (in which individuals can no longer be distinguished) if possible. If this is not possible, we strongly advise the use of adequate anonymisation techniques. While a full discussion of what exactly are 'adequate' anonymisation techniques is beyond the scope of this paper, we note that this can strongly vary from one type of data to another, especially in light of examples of data de-anonymisation by combining datasets (see e.g. Section 2.3).

Obtain permission from relevant stakeholders – Wherever possible researchers should seek permission from e.g. privacy and/or security officers for organisations where data is collected. Given that acquiring informed consent may sometimes be hard in the context of network measurement research, the people fulfilling these roles may be able to act as a proxy for their users. This is especially the case for privacy officers, so if this role is available in an organisation, talking to them is the preferred option.

The second impact to consider in case of passive measurements is the additional load the measurements may impose on production systems. In many cases, software for passive measurements will be run on the production systems that send and receive the traffic to be measured. The additional load this may impose can potentially interfere with the correct functioning of these production systems. This is undesirable, and to minimise the risk, we recommend taking into account the following guidelines:

Continuous performance monitoring – Most operators have monitoring in place to monitor the performance of their operational infrastructure. This type of monitoring can typically also be used to gauge the impact of measurements on the production systems. We recommend that researchers:

1. talk to operators to establish what is an acceptable upper limit to the additional load resulting from the measurement;
2. (if possible) set alarms on monitoring systems that trigger if this threshold is exceeded;
3. actively engage with operators to monitor the impact of the measurement (this also helps convey that researchers take an operator's concerns seriously and can help build a longer term relationship for future collaboration).

Offload filtering and analysis – Whenever possible, researchers should limit the software running on production infrastructure to just data capture, and to offload filtering and analysis of this data to systems that are not mission critical. We note that this recommendation may conflict with the desire to discard privacy-sensitive information as early as possible. This may require taking procedural measures to discard the privacy-sensitive data as soon after collection as possible, and may require explicit procedural guidelines in the data collection process.

4.2.3 Guidelines for active measurements

In case of active measurements our recommendations focus mostly on the impact that the measurement may have on the systems that are being probed or queried by the measurement. This does not mean that there are no privacy considerations for active measurements; if there is any reason to assume that personally identifiable information may be collected, researchers should also take the guidelines regarding privacy provided in the section on passive measurements into consideration.

Regarding the impact of the measurement, we provide six concrete guidelines listed below:

Re-use existing data and share results – Under the assumption that any active measurements impacts the systems that are studied in some form, the first step of any research project that is considering performing active measurements is to check for the existence of suitable datasets. Conversely, any research project for which active measurements are performed should strongly consider making its results public, or if that is not possible, accessible under restrictions. This allows other to do without additional measurements and re-use your data. A strong incentive for this practice may be to get exposure for your work. For example, a common practice when making data public is to require attribution, e.g. in the form of citing the paper(s) that relate to the data.

Clearly advertise intent – Active measurements are easily mistaken for scans with malicious intent, and can cost operators valuable time to investigate, if the intent of the measurement is not clearly communicated (as, e.g., the examples quoted in [23] illustrate). It is therefore of paramount importance that researchers clearly communicate the intent of their measurements in ways that are easy to find for operators. Three actions are recommended in particular:

1. Have a webpage with a clear description – This webpage should contain general information about your project and measurement(s), the intentions of your measurement and the type of traffic that operators can expect to see as a result of your measurement. In addition to this, their should be clear contact information for operators that want to reach you, for example to opt out (see below), or to report abuse. It is important to post regular updates on this webpage, to show that the project is being actively looked after.
2. Make measurements traceable – If you perform your measurements from specific IP prefixes, make sure these are clearly identified in the Regional Internet Registry's database, for example by adding a link to your project webpage in the description. Defining clear reverse DNS entries for measurement hosts is also important, for example something like scanner. visit www.myproject.org.
3. Communicate with incident response teams – It is key to inform local incident response teams that are responsible for the network(s) from which measurements are conducted. Inform them about when measurements will take place, expected traffic volumes, etc., and provide them with clear contact information so they can reach out if they receive abuse complaints.

Promptly Respond to Questions and Complaints – As stated above, it is important to make it easy for operators that have questions or complaints to reach you. It is equally important to respond promptly to any questions or complaints you receive directly, or through, e.g., your local incident response team. Ideally, a response should be sent within 24 hours, and we also recommend clearly advertising the time zone you are in on your webpage, to manage expectations on response times.

Provide Means to Opt Out – Any active measurement should incorporate an explicit means to opt out. Information about how to opt out should also be included in the information on the measurement project's web page.

Any request to opt out should be treated promptly, ideally in the same timely manner that abuse complaints are handled.

Actively Monitor Traffic Loads – Finally, we recommend actively monitoring the traffic load generated by the measurement, continuously if possible, but at least at regular intervals. If you expect your measurement to hit certain large operators more frequently, and you have the opportunity to set alerts on traffic peaks on your measurement infrastructure, we recommend that you do so.

4.3 Guidance for Third Party Data Suppliers

In many cases, network measurement research data is not collected directly by researchers, but is supplied to them by third parties, such as network operators. Broadly speaking, there are two ways in which third party data suppliers come into play:

- the third party shares data that it collects on a regular basis, independent of the research, e.g., for operational purposes;
- the third party collects data specifically at the request of the researcher.

In this first case, many of the issues and etiquette around the use of shared network data, identified by Allman et al. [10], will apply. Depending on the jurisdiction in which data is collected, specific legal requirements may apply, such as the General Data Protection Regulation (GDPR) of the European Union [25]. While it may be tempting to assume that if the data sharing complies with legal requirements, that is sufficient, we urge caution. Many legal frameworks centre around privacy concerns, and while this is often one of the key ethical concerns, it is certainly not the only one. Therefore, what is deemed legal, is not necessarily ethical. Therefore, we urge researchers and their IRBs to perform a thorough ethics review, even if data is supplied by a third party. Moreover, we encourage third parties to not just rely on IRB decisions, but to also establish their own ethical data sharing process, and provide a brief example case study of such a practice in Section 4.3.1.

The second case adds and additional challenge: what role should the third party data provider play in the ethics review? It is important that the third party actively participates in the ethics review process. They will have a unique perspective on the effects of the proposed measurement both on their own organisation, and on the users of their network or services on which data is collected. All of the considerations discussed above, that apply in case data that is regularly collected by the third party is shared, apply in this situation as well. In addition to this, however, we note that it is all the more important that the third party has their own ethical data sharing process in place. This allows the third party to assess independently whether the proposed research and associated data collection request aligns with their values.

4.3.1 SURFnet: a case study

As an example of third party ethical data sharing practices, we now present a brief case study. This case study centers on SURFnet, the National Research and Education Network in The Netherlands[6]. As discussed in Section 3, SURFnet actively participated in the Dagstuhl Seminar on Ethics in Data Sharing. After this seminar, SURFnet created a data sharing policy [20] that explicitly incorporates an ethics review. The policy was co-created with experts from the fields of ethics, law and computer network research. In broad outline, the policy requires that the following process is applied to data sharing requests to SURFnet:

Identify Data Risk Level – The prime concern that guided the creation of SURFnet's policy is that individual users or constituents of SURFnet[7] can be identified in the data. Therefore, the first step in the data sharing process is to gauge the risk for identification. The policy recognises three risk levels:

- *Low Risk* – Generally speaking, this aggregate, anonymous data that cannot feasibly be traced back to individual users.
- *Medium Risk* – Usually this is partially anonymised data, for example data to which a form of prefix-preserving IP anonymisation has been applied. Because this allows researchers to identify specific network segments in the data, it is classified in a higher risk category.
- *High Risk* – This is data that can be traced back to individual users (e.g., because original IP addresses are included, or because the data contains full packet captures).

Submit Ethics Section – SURFnet requires that researchers submit a brief section about the ethical considerations of their research. Most journals and conferences require researchers to do this as well, so SURFnet encourages

[6] https://www.surf.nl/en/about-surf/subsidiaries/surfnet/

[7] SURFnet's constituency consists of institutes of higher education and research institutes in The Netherlands.

researchers to write this part of their paper early on, so it can also serve the purpose of informing SURFnet about the ethical impact of the research.

Perform Ethics Review – In case research was classified as 'High Risk' in step 4.3.1, SURFnet requires researchers to participate in a mandatory ethics review. At its discretion, SURFnet may also require such a review for requests classified as 'Medium Risk'. In this ethics review, a specially convened ethics review board will study the ethical impact of the research. This review board will consist of three voting members, specifically the SURFnet employee who received the data sharing request, an independent SURFnet colleague not involved in handling the request and an external expert. The meeting of the review board is moderated by an external ethics adviser. Each meeting will conclude with a vote on the proposed research. The ethics adviser is a non-voting member of the board and will write an independent report about the meeting. If the ethics advisor does not agree with the decision reached by the review board, they can explicitly express this in the report.

Data Sharing Contract – In all cases, researchers will sign a data sharing contract with SURFnet. This contract stipulates the conditions under which the data is released, and if applicable (medium/high risk data), limits disclosure of the data. SURFnet recognises that academic best practices require data to be curated to enable reproducibile of research. At the same time, this may be at odds with non-disclosure requirements. In the case this clash occurs, SURFnet indicates in its policy that it will strive to offer long-term data curation for the research.

Provide Appropriate Attribution – As Allman et al. [10] already indicate, including appropriate attribution for the source of data used in research is important. As SURFnet is publicly funded, it has a moral obligation to facilitate network research. At the same time, in order to ensure it has sustainable funding, SURFnet needs to underline the benefits of its work to society. Academic research based on data shared by SURFnet is a prime example of societal benefit. Therefore, in its data sharing policy, SURFnet explicitly requires that researchers acknowledge SURFnet as the source of the data that was used for research.

For more information, SURFnet has made its data sharing policy publicly available on its website[8].

[8] https://www.surf.nl/datasharing

5 Conclusions and Recommendations

In this paper, we have studied the ethics around network measurement research from multiple angles. Our extensive overview of case studies about such research highlights systemic problems in how ethical considerations are taken into account by researchers and the institutions they belong to. An absence of sufficient guidance from Institutional Review Boards or Ethics Committees has led to serious ethical problems with network measurement-based studies. Despite efforts from the research community to establish clear guidelines (such as, e.g., the Menlo report [16]), there is ample evidence from recent studies (e.g. the censorship study discussed in Section 2.2) that there is still a long way to go.

In recent years, the discussion about ethical concerns regarding network measurement studies has surfaced at many different measurement conferences and workshops. While such discussions are a necessary first step to recognise that such concerns exist, it is of paramount importance that the community now takes the next step. Together, the network measurement research community should create an acceptable way of identifying and acting on ethical issues in measurement research. Such a practice must have broad support in the community. The risk is that if this does not happen, restrictions may be imposed from the outside. In this paper we presented a way forward for ethics reviews of network traffic measurement research, presented an outline on how to design ethically responsible measurements and also discussed an example best practice for third-party data suppliers. While these guidelines are in no way complete, they can serve as a starting point for a more coherent set of guidelines that is co-created by the Internet measurement community.

5.1 Recommendations

The case studies included in this paper show that IRBs and Ethical Committees sometimes forego advising on network measurement research because they perceive it as something that falls outside of their remit. None of the cases mentioned in this paper directly involve human subjects, only their data. Yet all of the cases can possibly have a strong impact on the human subjects related to that data. We strongly recommend that universities and research institutes, at which network measurement research takes place, explicitly include ethical reviews of this type of research into the mandate of their IRB or EC.

In addition to this, we strongly encourage that specific training programs are created to educate IRBs and ECs on the specific risks inherent in network measurement research. We envisage that educational materials to support this type of training can be created in collaboration between ethicists and the network measurement research community.

References

[1] Kramer, A. D. I., Guillory, J. E., and Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proc. Natl. Acad. Sci. U.S.A.* 111, 8788–8790.

[2] Burnett, S., and Feamster, N. (2015). Encore: Lightweight measurement of web censorship with cross-origin requests. *ACM SIGCOMM Comput. Commun. Rev.* 45, 653–667.

[3] Crandall, J. R., Crete-Nishihata, M., and Knockel, J. (2015). Forgive us our syns: Technical and ethical considerations for measuring internet filtering. *NS Ethics SIGCOMM*, 2015, 3.

[4] Harriman, S., and Patel, J. (2014). The ethics and editorial challenges of internet-based research. *BMC Med.* 12: 124. doi: 10.1186/s12916-014-0124-3

[5] Schroeder, C. (2015). Why can't we be friends: A proposal for universal ethical standards in human subject research. *J. Telecomm. High Tech.* 14, 409.

[6] Narayanan, A., and Zevenbergen, Y. B. No encore for encore? ethical questions for web-based censorship measurement. *SSRN Elect. J.* doi: 10.2139/ssrn.2665148

[7] Urban Development of Health and Health Services (2014). *Federal Policy for the Protection of Human Subjects (Common Rule)*. Available at: https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html

[8] Protection of human subjects. (2009). Available at: https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html

[9] Coull, S. E., Wright, C. V., Monrose, F., Collins, M. P., and Reiter, M. K. (2007). "Playing devils advocate: Inferring sensitive information from anonymized network traces," in *Proceedings of the Network and Distributed System Security Symposium*, San Diego, CA, 35–47.

[10] Allman, M., and Paxson, V. (2007). "Issues and etiquette concerning use of shared measurement data," in *Proceedings of ACM SIGCOMM IMC 2007*. San Diego, CA: ACM Press, 135–140.

[11] Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., and Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using facebook.com. *Soc. Netw.* 30, 330–342.

[12] Zimmer, M. (2010). But the data is already public: on the ethics of research in facebook. *Ethics Inf. Technol.* 12, 313–325.

[13] Poort, J., Leenheer, J., van der Ham, J., and Dumitru, C. (2014). Baywatch: Two approaches to measure the e_ects of blocking access to the pirate bay. *Telecommun. Pol.* 38, 383–392. 2014.

[14] van Wynsberghe, A., and van der Ham, J. (2015). Ethical considerations of using information obtained from online file sharing sites. *J. Inf. Commun, Ethics Soc.* 13, 256–267.

[15] van der Ham, J. (2015). Embedding ethics in system administration education. *USENIX J. Educ. Syst. Admin.* 2015, 1.

[16] Dittrich, D., and Kenneally, E. (2012). The menlo report: Ethical principles guiding information and communication technology research. *U.S. Depart. Home. Secur. Tech. Rep*. 2012, 514–516.

[17] Dittrich, D., Kenneally, E., and Bailey, M. (2013). Applying ethical principles to information and communication technology research: A companion to the menlo report. *U.S. Depart. Home. Secur. Tech. Rep*. 2013, 3.

[18] Partridge, C., and Allman, M. (2016). Ethical considerations in network measurement papers. *Commun. ACM*, 59, 58–64.

[19] Dietrich, S., Van Der Ham, J., Pras, A., van Rijswijk Deij, R., Shou, D., Sperotto, A., Van Wynsberghe, A. and Zuck, L. D. (2014). "Ethics in data sharing: developing a model for best practice," in *Proceedings of the Security and Privacy Workshops (SPW)*, IEEE. Rome, 5–9.

[20] van Rijswijk-Deij, R. (2015). "Ethics in Data Sharing: a best practice for NRENs," in *Proceedings of TNC 2015*. Porto.

[21] Bailey, M., Kenneally, E., and Dittrich, D. (2012). *A Refined Ethical Impact Assessment Tool and a Case Study of Its Application*. Berlin: Springer Berlin Heidelberg, 112–123.

[22] Cerf, V. G. (1991). RFC 1262 – Guidelines for Internet Measurement Activities. Available at: https://tools.ietf.org/html/rfc1262

[23] Papadopoulos, C., and Heidemann, J. (2009). "Towards best practices for active network measurement," in *Proceedings of the CAIDA AIMS Workshop,* San Diego, CA.

[24] van Rijswijk-Deij, R. (2017). *Improving DNS Security: A Measurement-Based Approach*. Ph.D. dissertation. University of Twente, Enschede.

[25] European Union (2016). *General data protection regulation 2016*. Available at: http://eur-lex.europa.eu/eli/reg/2016/679/oj1

Biographies

Jeroen van der Ham is a security researcher at the NCSC-NL since 2015. In his current research he focuses on privacy and security, as well as ethics in security research. He has published on ethical analyses of research and education, network monitoring, and semantic descriptions of computer networks and associated infrastructures. He currently holds positions at the TU Delft as well as the University of Amsterdam, where he serves as ethics advisor.

Jeroen received his Ph.D. degree from the University of Amsterdam in 2010 for his thesis entitled "A Complex Model for Computer Networks, the Network Description Language", after which he worked as a researcher at the University of Amsterdam until 2015.

Roland van Rijswijk-Deij works for SURFnet since 2008, and is a researcher at University of Twente. At SURFnet, the National Research and Education Network in The Netherlands, Roland is responsible for SURFnet's DNS and DNSSEC infrastructure. He also initiates and leads innovation projects in the area of Internet security and stability. Past innovation projects initiated by Roland have focused on DNS, DNSSEC, detecting and mitigating DDoS attacks, IPv6 and many other topics. Roland regularly presents his work in international networking venues, such as TNC, Internet2 conferences, IETF meetings, ICANN meetings, RIPE meetings and NANOG.

Roland received a cum laude Ph.D. degree from the University of Twente in June 2017, for his thesis entitled "Improving DNS Security: a Measurement-Based Approach".