Conducting an Ethical Study of Web Traffic

L Jean Camp
ljcamp@indiana.edu

John Duncan
johfdunc@indiana.edu

School of Informatics and Computing, Indiana University, Bloomington, IN, USA

Abstract

We conducted a study of student web browsing habits at Indiana University's Bloomington campus, in which we examined the web page requests of over 1,000 students during a period of two months. In this paper we present the ethical issues of this study, as opposed to the results. Specifically, we discuss the details of the study development and implementation from the point of view of ethical design, a framework cased on value-sensitive design. Concerns with stakeholder privacy, the quality of study data collection, human subjects research protocols, and unexpected data anomalies are presented in order to illustrate the many difficulties and ethical pitfalls confronting network researchers. Places in which the study met, and those where it failed, the principles of ethical design are highlighted. A secondary contribution is the evolution of the instruments which were developed through the human subjects process.

1 – Introduction

We sought to evaluate the inadvertent wisdom of crowds in avoiding dangerous web locales. Similarly to the fact that people do not visit dangerous areas; and therefore people do not loiter in urban areas bereft of humans; then if it visible that a website is never visited If users exhibit similarities in browsing habits (e.g. visit the same sites), they can better collaborate to combat online threats such as phishing. In order to establish and test metrics for user browsing homophily and session behavior, we conducted a study of student web browsing habits in an academic dormitory at Indiana University's Bloomington campus. Under the aegis of this study, we examined the web page requests of over 1,000 students during a period of two months. The results were informative (Meiss et al., 2009). However, the process of conducting such a study ethically is also academically relevant and important, and for this reason, in this paper we discuss the details of the study development and implementation. Ethical systems are ones that respect the rights of all stakeholders in the use of that system, whose development process places an emphasis on user participation, as well as feedback in the design of the system, and that value user satisfaction as highly as meeting functional requirements (Friedman, 1996; Friedman & Nissenbaum, 1996; Nissenbaum, 1998a, 1998b). The increasing concern with ethical design in the engineering of physical systems suggests that ethical design for software systems and research protocols is also increasingly important. (Cummings, 2006) This is especially notable in network research, where as the Director for CAIDA (Cooperative Association for Internet Data Analysis) notes, "No one is investing in technology to learn about networks while minimizing the amount of privacy compromised in the process" (Claffy, 2008).

Working in tandem with Indiana University's Human Subjects Committee (HSC) and Institutional Review Board (IRB), we designed our study in such a way as to minimize student impact. Our secondary goal was providing education by informing students as to the possible ways in which they might make their web browsing information difficult to detect by malicious snooping agents. To do this, we provided the students with information regarding the ways in which they might opt out of the study (while also increasing their security) through the use of a VPN. While we had desired to also provide students with information about systems for secure anonymous browsing, such as Tor (Dingledine, Matthewson, & Syverson, 2004), we did not receive approval to distribute this information.

To implement this study, we were first required to complete an exam on how to conduct human subjects research (Indiana University, 2011). This was geared primarily towards researchers who would be conducting face-to-face interviews or handing out physical questionnaires. Details on the questions asked by Human Subjects, and our responses to them, can be found in Section 7. Fliers were printed and distributed to the mailbox of every student involved in the study, several weeks before the study began. The progression of these fliers, from initial specification to final version, can be found in Appendix A. This progression illustrates the changes that were made to comply with human subjects protocols and to clarify the nature of the study to the subjects. From these fliers we received a small amount of student inquiries, and a request to coordinate with student government. After some

L. Jean Camp & J. Duncan, "Conducting an Ethical Study of Web Traffic", 5th Workshop on Cyber Security Experimentation and Test, A USENIX Security Symposium Workshop, (Bellevue, WA) 6 August 2012.

analysis and discussion with the HSC, it was determined that student government was in fact not a stakeholder in this study, as oversight in student research falls to the IRB and HSC, who are specifically charged to safeguard such participants. Overall, the small response suggested that students demonstrated a general lack of concern with the operation of the study. Finally, we also include the contents of an addendum to the study we filed (Appendix B), indicating the care which was taken with student data.

In addition to these measures, we also proposed to arrange a workshop for students in the selected dormitory, in which we would also cover several popular and useful tools for computer security. These tools ranged from anonymous browsing software to malware detection software and browser plugins designed to provide users with more control over the execution of scripting languages. Included in this proposal was a short talk covering basic principles of security, and other things students might do to protect themselves, such as update anti-virus and anti-malware software frequently and automatically. Ultimately, this workshop was rejected by the selected dorm's administration, who did not feel it to be of sufficient value to the students. Overall, dialog with student stakeholders was difficult due to the restrictions on contacting them.

2 - Technical Design and Anonymization Methods

The dormitory study was implemented by installing a dedicated FreeBSD server located in the central routing facility of the Bloomington campus of Indiana University. This system had a 1 Gbps Ethernet port that received a mirror of all outbound network traffic from one of the undergraduate dormitories. This dormitory consists of four wings of five floors each and is home to just over a thousand undergraduates. Its population is split roughly evenly between men and women, and its location causes it to have a somewhat greater proportion of music and education students than other campus housing.

To obtain information on individual HTTP requests passing over this interface, we used a Berkeley Packet Filter to capture only packets destined for TCP port 80 (the normal port for web browsing). These packets were mirrored to this traffic monitor, which examined the requested URL to determine if it was a web page (.htm, .html, .asp, .php, etc.) or other content (images, video, etc.). This methodology was also employed in (Gribble, 1997). We used a regular expression search against the payload of the packet to determine whether it contains an HTTP GET request. We then analyzed the packet to determine the identity of the virtual host contacted, the path requested, the referring URL, and the advertised identity of the user agent. We then wrote a record to our raw data files that contained a timestamp, the virtual host, the path requested, the referring URL, and a flag indicating whether the user agent matches a mainstream browser (Internet Explorer, Mozilla/Firefox, Safari, or Opera). The aggregate traffic of the dormitory was sufficiently low so that the sniffing system was able to maintain a full rate of collection without dropping packets. No content was fetched by the packet monitor during the study. No attempts to observe encrypted traffic were made (nor did we capture requests to TCP port 443), although classic man-in-the-middle attacks would have made such observation possible.

The click data was collected over a period of about two months, from March 5, 2008 through May 3, 2008. This period included a week-long vacation during which no students were present in the building. During the full data collection period, we logged nearly 408 million HTTP requests from a total of 1,083 unique MAC addresses. We filtered out a small subset of users with negligible activity; their traffic consisted largely of automated Windows Update requests and did not provide meaningful data about user activity. We also found that some web clients issue duplicate HTTP requests (same referring URL and same target URL) in nearly simultaneous bursts. These bursts occur independently of the type of URL being requested and are less than a single second wide. We conjecture that they may involve checking for updated content, but we are unable to confirm this without access to the original HTTP headers. Because this behavior is so rapid that it cannot reflect deliberate activity of individual users, we also removed the duplicate requests from the data set. The resulting data set is the basis for all of the discussion in this paper. It contains a total of 29.8 million page requests from 967 unique users. These requests were directed toward a total of over 630,000 distinct Web servers, and nearly 110,000 distinct servers provided referrers for those requests.

3– Related Work

Students, as a population being considered for an application, are stakeholders. Even when participating as an anonymized aggregate, stakeholders' rights must be respected. Derived from Value-sensitive design, our process for ethical design informed the framing of the study and definitions of data. Risks, in this study, took the form of incidents or data capture policies that would threaten subject privacy. Identifying such risks and eliminating them

when possible, while preserving the functionality of the data, was a major component of the study. Where we were unable to uphold the standards of ethical design completely, we discuss what occurred and identify points in which further work is needed. By beginning a dialogue on ethical study design in network research, we hope to strengthen privacy and security for the studied populations, while preserving the ability for researchers to capture the data needed for their work. This potential tension between the needs of stakeholders, researchers, and research review boards is identified by Mosher as a "justice-as-fairness" concern (Mosher, 1988). While he identifies this tension as most visible in the area of controversial human-subjects research on subjects such as sexuality, the controversial nature of traffic studies on real-world networks, involving actual usage data, renders his arguments equally applicable.

During our data collection, requests for URLs were trimmed at the file extension, to prevent the inclusion of any session-based information (such as cgi variables) that would identify the user. Cockburn et al. performed similar trimming, although theirs was primarily designed to correctly register multiple visits to sites that employ search parameters (Cockburn, McKenzie, 2001). When these trimmed URLs were added to the database of browser requests, we attached an identifier constructed by taking the MAC address of the requesting party and passing it through a suitable one-way function to produce a pseudo-anonymous identifier. By choosing the MAC address as the origin of our identifiers, we made the assumption that most computers in the building have a single primary user.

Removing session identifiers and passing the MAC address through a one-way function are often characterized as sufficient to preserve anonymity. A similar method was used in Bestavros et al., although no analysis of its effectiveness was performed (Bestavros et al., 1995). Gribble employed hash functions more extensively - to conceal source IP, destination IP, and requested URL (Gribble, 1997). From the pattern of their requests, it may or may not be possible to identify users. This increases in likelihood if user traffic is highly dissimilar (i.e. exhibits low homophily). If it is not possible to infer identity directly from traffic patterns, it may be possible to create a profile that allows tracking even without current methods such as tracking cookies. Lasko and Vinterbo suggest that removing identifiers and obfuscating the MAC address are not necessarily sufficient to prevent users from being identified. They describe a tension between privacy protection and analytic utility, suggesting that techniques such as their proposed spectral anonymization must be used to maintain subject privacy in high-dimensional datasets. This technique was not publicized at the time of our study, however, it appears to offer some further subject protections (Lasko, Vinterbo, 2010).

While our data was removed post-analysis, and by the very design of the study, it was never intended to be shared, there is some argument for the benefit to all stakeholders that might arise from increased data sharing among researchers. Kenneally and Claffy, specifically, argue that data sets such as ours, with first-order and second-order identifiers removed are less likely to pose privacy risks for subjects (Kenneally, Claffy, 2010). Where researchers have defaulted to not sharing data sets, this is often due to unclear policies on data sharing permissibility, and less on an inherent concern for stakeholder privacy.

4– Difficulties in Ethical Research

In the course of the data analysis for this study, we also discovered the presence of a poorly-written anonymization service that was attempting to obscure traffic to an adult chat site by spoofing requests from hundreds of uninvolved clients. Adjusting the technical parameters of the study was a simple matter: these requests were removed from the data set because they were not genuine. However, this behavior was identified with enough time left in the study to have altered the code collecting our data to retrieve and store the actual MAC address of the interface card making the requests. From there, with the assistance of university technology professionals, we could have located the actual student machine responsible, and presumably, its owner. In doing so we might have informed this student of the poorly-coded nature of this software and directed them to specific other anonymization resources, such as Tor, that provide more robust anonymity (Dingledine, Matthewson, 2006).

Ultimately, we decided not to make an effort to contact the student responsible for these anomalies. Simply put, we would have had to violate the carefully delineated boundaries of the study to do so, and may have caused the student embarrassment even if they were contacted, say, by email. The chance that they might avoid future embarrassment from the detection of their behavior simply did not outweigh the immediate costs. Had the student been using this anonymization method to view non-controversial material, it would have been more likely that we would have contacted them. No matter how carefully delineated a policy for data collection is, there can always be a

user that creates a perfect window of uncertainty. This incident is an excellent example of the tension between security, privacy and trust issues that may arise while conducting ethical research. Had we discovered a user who was visiting child pornography sites, we might have been legally obligated to report this occurrence. Steinberg et al. identify this concern in the context of child abuse and clinical studies, but this dilemma is equally applicable in network research. The central question, as they put it, is "whether researchers have a moral duty to place the health and safety of children above concerns about confidentiality and the benefits of obtaining new knowledge" (Steinberg et. al, 1999). Solutions to this problem require further exploration of the ethical responsibilities of researchers who are conducting studies in which data is being anonymized in real-time.

5– Failures to Comply with Ethical Design

In several situations, the execution of this study was not able to wholly comply with the philosophy of ethical design. To begin with, the physical structure of the network architecture imposed severe limits on our ability to conduct the study with an opt-in recruitment model, which more fully preserves the rights of stakeholders. While contacting individual students was possible, and it is certainly possible that these students might have been able to successfully identify and transmit their MAC addresses to us for the purposes of the study, the packet monitor could only have been attached at a central data collection point. There, each packet could have been compared to a list of approved MAC addresses, but this would have involved viewing unauthorized packets simply for the purposes of determining whether or not they were originated by valid participants, thus violating the opt-in model entirely. The assumption of one student to one MAC address would also need to be made in this case. As well, during particularly high traffic times this approach risks discarding packets by adding several additional processing steps. Finally, concern about the number of students who would be willing to undertake the steps necessary to identify a MAC address and provide it prompted us to decide that the study would be much more accurate with a larger population. In this case, technological limits drove study choices, even when ethical design advocated different methods.

While notifications were provided to the students by mail, we had no way of accurately assessing how many of them read or understood the notices, outside of the small number of emailed questions which were received about the study. Generally, the students appeared to exhibit a lack of concern, but this is not the same as full understanding and consent. Similarly, while a method to opt out was provided, opting out is fundamentally less friendly to a study population than opting in, whether or not the method of opting out increases user security. Since we were not, in the end, permitted to conduct an educational panel at the dormitory, we cannot be sure what student response or understanding of the study truly was.

6– Successes with Ethical Design

While the above section details the difficulties we had in conducting this study in a manner fully in accord with ethical design, there were also several notable areas in which we succeeded in doing so. During design, data collection, and analysis, when the study generated or identified risk all data were purged of identifiers or other elements of identifiable risk to the subjects, whether or not this was initially in accordance with our design methodology.

Because opt-in was not technologically friendly nor likely to generate sufficient data for the study, we took careful steps to ensure that no identifiable data was recorded on any individual, in accordance with current best practices for aggregate data anonymization. The data set was retained only for the period of the data analysis. In the two cases in which stakeholders contacted us with concerns, we communicated with each of them to discuss how these concerns might be mitigated in a technological and procedural context. Aside from the subject who had concerns about student government involvement, the other student had a simple question about study dates. Limited student inquiries make it difficult to assess how successful this approach was in informing students of their rights and any potential risks, as we could not conduct any follow-up studies with individuals.

7– Experiment Proposal & HSC Review

In this section, we present a selection of the questions asked by the Human Subjects Committee, and the answers we gave to them, in order to illustrate the concerns that were addressed in the course of our receiving approval to proceed. Questions are presented inset in smaller font, and our answers follow. While this framework is geared towards research performed in a laboratory setting, it still assisted us somewhat in our study design.

Briefly describe, in lay terms, the general nature and purpose of the proposed research, and where the study will take place. If student research, indicate whether for a course, thesis, dissertation, or independent research. If the study is only for a course, please review the Student Research Policy to ascertain if this project requires HSC review.

The purpose of this study is to measure the degree of homophily (same-ness) in web browsing traffic, and the effectiveness of current standard practice for de-identifying individuals. The first part (the degree of same-ness) is important because it governs the degree to which users in social networks can gain a benefit by voluntarily cooperating to protect themselves against hostile agents. The second part (current practices for anonymization) is important, because the degree to which these practices are effective has not been properly evaluated. This research will allow us to better protect users from malicious agents on the web.

To implement this study, we plan to record certain parts of the requests for web pages made by web browsers. In particular, we plan to record the URLs in requests made by students in an IU Bloomington residence hall. By following currently accepted anonymization protocols, it is believed that no individual student will be identifiable during this process. However, testing this assumption is also part of this research. The anonymized data will be collected by a secure server, and then placed onto a CD for processing. This CD will be stored in a locked room.

Specifically, we will be recording traffic to port 80 (the port web browsers use to make their requests). Non-browser traffic to this port will be ignored. We will not actually be fetching the web pages requested, or any content thereof. Any traffic that is secured (i.e. encrypted or otherwise secure traffic, such as email, bank traffic, or records) cannot be monitored by our system and will not be recorded. URLs (web page addresses) that contain any sensitive information that could be used to identify individuals will have this information removed prior to storage. To monitor this information, special hardware will need to be installed in the network closets of the residence hall. We have been coordinating this request with UITS.

Describe the process by which subjects will be recruited, how many (or estimate) subjects will be involved in the research, and how much time will be required of them. List specific eligibility requirements for subjects (or describe screening procedures), including those criteria that would exclude otherwise acceptable subjects. If your study uses only male or female subjects, explain why. For NIH-funded research only, address the inclusion of women, minorities and children in the research. Disclose any relationship between researcher and subjects - such as, teacher/student; superintendent/principal/teacher; employer/employee.

A particular residence hall on the IU Bloomington campus will be selected based on its technical suitability for this research, as decided by UITS. No monetary gain or reward will be offered for this study.. No investment of time or resources is required by the users. There is no relationship between the researchers and the subjects. No particular sub-population of students is being targeted for this research.

We plan to collect 2 months worth of traffic data. As a student holiday such as Spring Break may fall during this period, and other delays may be involved in deploying the system, we would like to request 10 weeks to perform the data collection period of this study.

List all procedures to be used on human subjects or describe what subjects will do. If done during regular

> class time, explain what non-participants will do. If you are taping, explain that here (see item 13 on page 11). *Asterisk* those you consider experimental. For those asterisked procedures, describe the usual method(s), if any, that were considered and why they were not used. (See item F on page 2 for more information.)

Subjects will simply continue to use web browsers as normal. Any page requests they make that are unencrypted will have their URLs recorded and be anonymized relative to the user, according to current best practices.

> Describe methods for preserving confidentiality. How will data be recorded and stored, with or without identifiers? If identifiers are used describe the type: names, job titles, number code, etc. How long are identifiers kept? If coding system is used, is there a link back to the subject's ID? If yes, where is the code list stored in relation to data and when is the code list destroyed? How will reports will be written, in aggregate terms, or will individual responses be described? Will subjects be identified in reports (see item 5 on page 10)? Describe disposition of tapes/films at the end of the study. If tapes are to be kept, indicate for how long and describe future uses of tapes.

We will be using a one-way function to convert the hardware address of each student's computer (an identifier) to a number in a manner that is non-reversible. This number will be the only tag attached to the page requests we record. This is currently believed to be sufficient for anonymity, and there is no link back to the hardware address.

Reports about the data gathered will be presented in aggregate form – no individual history will be presented for any user. While it is believed that no subject can be identified from our data, if we should succeed in doing so, this result will not be presented in conjunction with any individual identity. After the study, data will be retained only on a single encrypted CD stored in Dr. Camp's room in a locked cabinet. This CD will not be made publicly available.

We will attempt to remove all information in the truncated URLs that could be used to identify individuals prior to storage in our research dataset, and will be keeping track of the number and kind of these instances (without any personally identifiable information associated with those instances) as an important research finding. In order to perform this removal, we will need to personally examine the dataset, as no automated process exists to perform this operation.

> What, if any, benefit is to be gained by the subject? In the event of monetary gain, include all payment arrangements (amount of payment and the proposed method of disbursement), including reimbursement of expenses. If class credit will be given, list the amount and the value as it relates to the total points needed for an A. List alternative ways to earn the same amount of credit. If merchandise or a service is given, indicate the value. Explain the amount of partial payment/class credit if the subject withdraws prior to completion of the study. (See policy at http://www.research.indiana.edu/rschcomp/compensation.html)

No individual rewards will be offered to subjects. However, users will be presented in the attached notice with easy steps to take to increase their security while browsing the web. Even after the study has been completed, if they continue to follow these steps they will continue to receive increased security.

> What information may accrue to science or society in general as a result of this work?

Data from this study will help lead to better systems for protecting users from malicious agents

on the web. Additionally, our research about current anonymization techniques will help establish whether these methods, commonly used today, are actually accomplishing their goals. If not, new techniques need to be developed to protect user privacy. This finding will be extremely important either way.

Is modification of the required elements for informed consent requested?

For technical reasons, as discovered by UITS, it is not feasible to obtain individual consent from each potential subject. Instead, we are attaching a notice, which would be disseminated in hardcopy to every student living in the targeted residence hall. This notice outlines the purpose and practice of the research, and presents ways to opt out of it (which also increase each individual student's security), as well as the information that no penalty or harm comes from opting out.

Explain how this research involves no more than minimal risk to the subject. Loss of confidentiality is, under most circumstances, more than minimal risk. However, contact by primary care givers or others who by the nature of their involvement with the subject already have access to the data, will be considered no further loss of confidentiality and, therefore, may be less of a risk to subject confidentiality. Risk may also vary with the type of information being collected.

Users are being de-indentified according to current best practices for doing so. As well, all data collected is non-encrypted, and is not technically protected from eavesdropping. Confidential information such as web browsing in regards to email, financial services, or academic records is protected by encryption and thus explicitly non-recordable for this research.

Explain how the research could not practically be carried out without waiver of informed consent.

Technically, it is not possible to opt-in subjects for this experiment. Because of the way that networking functions, an area is targeted. For this reason, we are presenting students with a way to opt-out instead, which is technically feasible. No deception is involved in this research.

Explain how, if appropriate, subjects will be informed of pertinent information after participation.

Students will be informed of all pertinent information by the attached notice. Students will receive the notice as a flyer in their mailbox. We will also be working with Dorm management to disseminate this information through any channels they are willing to make available to us.

8– Conclusion

Ethical systems should not impose technological fiat on their users. Due to technological restrictions, we were forced to use an opt-out model, rather than an opt-in model. Users who wished to remove themselves from the study were required to take specific steps to do so. While these steps increased participant network security, it is still not desirable to force users to take action to avoid participation. Ethical systems should take steps to reduce data footprint whenever possible, especially when the data involved belongs to stakeholders or derives from stakeholder actions. While the collection of network traffic data at the packet level will never generate a small data footprint, we held true to the principles of ethical design by observing constraints on data granularity. Pseudonymity was used to protect subjects, and only minimal packet data were stored. No content was retrieved. Ethical systems address

incentive misalignment issues should these emerge, engaging all stakeholders in this process. This study was designed to meet the needs of researchers, but subjects did not experience any difference in their normal activities if they chose to remain in the study. Ethical systems allow stakeholders involvement at all levels of the design process. Limited means of contacting subjects and a general lack of stakeholder concern made it difficult to fully involve users at the desired level.

Ultimately, this study allowed us to experience the difficulties than can emerge from attempting to comply with ethical design principles when technological constraints and limited stakeholder response make it difficult. This research shows examples of success and failure in compliance with our initial ethical goals.

9 – Special Acknowledgements

We would like to thank the various agencies and individuals who made this study a possibility, including Mary Beth Schmucker, Robert Weith, Residential Programs and Services (RPS), University Information Technology Services (UITS), the Human Subjects Committee (HSC), the Institutional Review Board (IRB) and the staff in Residential Services. At each step in the process of this study we had assistance from one or more of these groups and individuals, without whom the study might never have been conducted at all.

10 – References

Bestavros, A., Carter, R., Crovella, M., Cunha, C., Abdelsalam, H., & Mirdad, S. (1995). Application-level Document Caching in the Internet. *IEEE SDNE.*

Claffy, KC (2008). Ten Things Lawyers Should Know About the Internet. Retrieved March 20, 20011, from CAIDA: http://www.caida.org/publications/papers/2008/lawyers_top_ten/

Cockburn, A., & McKenzie, B. (2001). What Do Web Users Do? An Empirical Analysis of Web Use. *International Journal of Human-Computer Studies, Vol 54, Issue 6* .

Cummings, M. L. (2006). Integrating Ethics in Design through the Value-Sensitive Design Approach. *Science & Engineering Ethics*, 12(4), 701-715.

Dingledine, R., Matthewson, N., & Syverson, P. (2004). Tor: The Second-Generation Onion Router. *Usenix Security, August.*

Dingledine, R., & Matthewson, N. (2006). Anonymity Loves Company. *Workshop on the Economics of Information Security, June.*

Friedman, B. (1996). Value-Sensitive Design. *Interactions Volume 3* , 16-23.

Friedman, B., & Nissenbaum, H. (1996). Bias in Computer Systems. *ACM Transactions on Information Systems Vol 14* , 330-347.

Gribble, S. (1997). System Design Issues for Internet Middleware Services: Deductions form a Large Client Trace. *Proceedings of the Usenix Symposium on Internet Technologies and Systems*, (pp. 207-218).

Indiana University. (2011, February 26). *IU: Office of Research Administration*. Retrieved February 26, 2011, from IU ORA: http://researchadmin.iu.edu/HumanSubjects/

Kenneally, E., & Claffy, KC (2010). Dialing Privacy and Utility: A Proposed Data-Sharing Framework to Advance Internet Research. *IEEE Security and Privacy*.

Lasko, T. A., & Vinterbo, S. A. (2010). Spectral Anonymization of Data. *IEEE Transactions on Knowledge & Data Engineering*, 22(3), 437-446.

Meiss, M., Duncan, J., Goncalves, B., Ramasco, J., & Menczer, F. (2009). What's in a Session: Tracking Individual Behavior on the Web. *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia (HT 2009).*

Mosher, D. L. (1988). Balancing the Rights of Subjects, Scientists, and Society: 10 Principles for Human Subject Committees. *Journal of Sex Research*, 24(1-4), 378.

Nissenbaum, H. (1998a). *Values in Computer System Design: Bias and Autonomy.* Delhi: New Academic Press.

Nissenbaum, H. (1998b). Values in the Design of Computer Systems. *Computers in Society* , 38-39.

Steinberg, A. M., Pynoos, R. S., Goenjian, A. K., Sossanabadi, H., & Sherr, L. (1999). Are Researchers Bound by Child Abuse Reporting Laws?. *Child Abuse & Neglect*, 23(8), 771-777.

Appendix A – Experiment Notices

Six versions of the experiment notice are presented here, covering the initial draft of the notice to the version that was ultimately approved and distributed to students. Changes and additions are highlighted in order to illustrate the full process by which we arrived at a consensus between the research team, the IRB, and the Human Subjects Committee. Specific changes included the alteration of phrases that were viewed as potentially unclear to students, such as "publically available data" to "a truncated version of the URLs". This allowed students to better understand the scope of the research, and to decide whether or not they wished to opt out.

A.1 – First Version – 11/13/2007

Notice – Your residence hall has been selected to participate in an IU research study!

A research team here at IU has designed this study to measure two things. First, to what degree does the similarity of a group of people relate to the similarity of their web browsing habits? Second, how well do traditional methods to protect privacy (such as anonymizing web traffic) truly protect private information?

Studying these things allows us to better understand the value of various existing privacy-enhancing technologies, such as Tor (http://www.torproject.org/). We also want to better understand if a group of people in a social network (like a friends group on MySpace or Facebook) can protect themselves from certain types of hostile web sites (fraudulent commerce sites, web sites with viruses, etc.) by voluntarily sharing information. This research relates directly to the Net Trust project, developed here at IU (http://www.ljean.com/NetTrust/).

In order to conduct this experiment, we will be recording publicly available data about web browser requests made in this residence hall. We will not record: your network id, the content of requested pages, passwords, usernames, or any other content information – only the address of the page requested. Non-public browser requests (sites that use encryption such as your email, financial institutions, academic records, and Oncourse) will not be viewed or recorded in this experiment.

If you decide you don't want to participate in this experiment, several tools are available that convert all public information transmitted by your web browser into private data, which will then not be recorded in this experiment. Tor (http://www.torproject.org/) is one of these tools, as is IU's Virtual Private Network. An installer for the VPN can be obtained from IUWare (http://iuware.iu.edu/list.aspx?id=134). These links also contain information about how to install and use these tools. Students who decide not to participate in this research will not be penalized in any way for doing so, nor can you be identified as someone who opted out.

This experiment will take place from xx/xx/0x to yy/yy/0y.

For more information on this experiment or the research itself, please contact the Primary Investigator, John Duncan, at johfdunc@indiana.edu . Please put "Browser Study" in your Subject line.

Thanks for supporting IU Research!

A.2 – Second Version – 11/16/2007

Notice – Your residence hall has been selected to participate in an IU research study!

A research team here at IU has designed this study to measure two things. First do similar people (geographically and demographically) have similar web browsing habits? Second, how well do traditional methods to protect privacy (such as anonymizing web traffic) truly protect private information?

Studying these things allows us to better understand the value of various existing privacy-enhancing technologies, such as Tor (http://www.torproject.org/). We also want to better understand if a group of people in a social network (like a friends group on MySpace or Facebook) can protect themselves from certain types of hostile web sites (fraudulent commerce sites, web sites with viruses, etc.) by voluntarily sharing information. This research relates directly to the Net Trust project, developed by the same researchers who are doing this study (http://www.ljean.com/NetTrust/).

In order to conduct this experiment, we will be recording publicly available data about web browser requests made in this residence hall. We will not record: the content of requested pages, passwords, usernames, or any other content information – only the address of the page requested. Data transmitted during non-public browser requests (sites that use encryption such as your email, financial institutions, academic records, and Oncourse) can not be recorded in this experiment.

If you decide you don't want to participate in this experiment, several tools are available that convert all public information transmitted by your web browser into private data. Remember that private data will not be recorded in this experiment. Tor (http://www.torproject.org/) is one of these tools, as is IU's Virtual Private Network. An installer for the VPN can be obtained from IUWare here: http://iuware.iu.edu/list.aspx?id=134 . These links also contain information about how to install and use these tools. Students who decide not to participate in this research will not be penalized in any way for doing so, nor can you even be identified as someone who opted out.

This experiment will take place from xx/xx/0x to yy/yy/0y.

For more information on this experiment or the research itself, please contact the Primary Investigator, John Duncan, at johfdunc@indiana.edu . Please put "Browser Study" in your Subject line.

Thanks for supporting IU Research!

A.3 – Third Version – 11/26/2007

Notice – Your residence hall has been selected to participate in an IU research study!

A research team here at IU has designed this study to measure two things. First do similar people (geographically and demographically) have similar web browsing habits? Second, how well do traditional methods to protect privacy (such as anonymizing web traffic) truly protect private information?

Studying these things allows us to better understand the value of various existing privacy-enhancing technologies. We also want to better understand if a group of people in a social network (like a friends group on MySpace or Facebook) can protect themselves from certain types of hostile web sites (fraudulent commerce sites, web sites with viruses, etc.) by voluntarily sharing information. This research relates directly to the Net Trust project, developed by the same researchers who are doing this study (http://www.ljean.com/NetTrust/).

In order to conduct this experiment, we will be recording publicly available data about web browser requests made in this residence hall. We will not record: the content of requested pages, passwords, usernames, or any other content information – only the address of the page requested. Data transmitted during non-public browser requests (sites that use encryption such as your email, financial institutions, academic records, and Oncourse) can not be recorded in this experiment.

If you decide you don't want to participate in this experiment, several tools are available that convert all public information transmitted by your web browser into private data. Remember that private data will not be recorded in this experiment. IU's Virtual Private Network is an easy way to improve your security while removing yourself from this study. An installer for the VPN can be obtained from IUWare here: http://iuware.iu.edu/list.aspx?id=134 . Students who decide not to participate in this research will not be penalized in any way for doing so, nor can you even be identified as someone who opted out.

This experiment will take place from xx/xx/0x to yy/yy/0y.

For more information on this experiment or the research itself, please contact the Primary Investigator, John Duncan, at johfdunc@indiana.edu . Please put "Browser Study" in your Subject line.

Thanks for supporting IU Research!

A.4 – Fourth Version 01/18/2008

Notice: Your residence hall has been selected to participate in an IU research study.

We are giving you this notice of our research study so that you may choose whether or not to participate. See below for a description of the study and what you need to do if you do not wish to participate.

Who are we and what are we studying?

Our research team at the IU School of Informatics is working to better understand what makes people vulnerable to fraud and other scams when using the Internet, and what types of solutions may help protect them against such threats. As part of these efforts, we have designed a study to measure two things. First, do people who are similar geographically and demographically have similar web browsing habits? Second, how well do traditional methods used to protect privacy truly protect private information? Studying these things will allow us to better assess existing technologies for protecting privacy, and whether a group of people in a social network can protect themselves from certain types of harmful web sites by voluntarily sharing information.

What information are we collecting as part of this study?

We will be recording a truncated version of the URLs (web addresses) of the web pages that residents request as the requests are transmitted over the IU network from your dorm to the Internet. We have chosen to study web browsing from a single dorm because the residents are by definition geographically similar, and they also tend to be similar demographically.

What information are we NOT collecting as part of this study?

We will NOT record any of the following:

- names, demographic information, or other personally identifiable information.
- the full URL or any content of web pages any resident browses. We will only record a truncated URL. We do not expect to encounter any personally identifiable information in the truncated URLs we are collecting, but if we do, we will delete it promptly.

We will NOT know the identity of a resident browsing the Internet and will NOT be viewing the content of the web pages a resident visits.

What do I need to do if I don't want to participate in the study?

Any web traffic that is encrypted – i.e., traffic that is sent via https:// or through a VPN connection – cannot be captured or recorded as part of this study. So if you decide you don't want any of your web browsing information to be recorded as part of this study, you can encrypt your web traffic in one of two ways:

- ensure that the website you are using has a URL starting with https:// instead of http://
- use IU's Virtual Private Network (VPN) service to do all of your web browsing. For more information about how to use the VPN, see http://kb.iu.edu/data/ajrq.html, or call the UITS Support Center at x5-6789.

There is no penalty for not participating in this study. Since no personal information will be recorded, we won't be able to tell whether a given resident even participated or not.

When will the study take place?

We will record information between [dates]

Has this study been approved by Indiana University?

Yes. The IU Bloomington Human Subjects Committee and Residential Programs and Services have both approved this study.

If I have questions about the study, who should I call? For more information on this study or its procedures, please contact the Primary Investigator, John Duncan, at johfdunc@indiana.edu. [phone number removed]. If you have any concerns about the study or its procedures, please contact the office for the Indiana University Bloomington Human Subjects Committee, Carmichael Center L03, 530 E. Kirkwood Ave., Bloomington, IN 47408, [phone number removed], by email at iub_hsc@indiana.edu

A.5 – Fifth Version – 01/26/2008

Notice: Your residence hall has been selected to participate in an IU research study.

We are giving you this notice of our research study so that you may choose whether or not to participate. See below for a description of the study and what you need to do if you do not wish to participate.

Who are we and what are we studying?

Our research team at the IU School of Informatics is working to better understand what makes people vulnerable to fraud and other scams when using the Internet, and what types of solutions may help protect them against such threats. As part of these efforts, we have designed a study to measure two things. First, do people who are similar geographically and demographically have similar web browsing habits? Second, how well do traditional methods used to protect privacy truly protect private information? Studying these things will allow us to better assess existing technologies for protecting privacy, and whether a group of people in a social network can protect themselves from certain types of harmful web sites by voluntarily sharing information.

What information are we collecting as part of this study?

We will be recording a truncated version of the URLs (web addresses) of the web pages that residents request as the requests are transmitted over the IU network from your dorm to the Internet. We have chosen to study web browsing from a single dorm because the residents are by definition geographically similar, and they also tend to be similar demographically.

What information are we NOT collecting as part of this study?

We will NOT record any of the following:

- names, demographic information, or other personally identifiable information.
- the full URL or any content of web pages any resident browses. We will only record a truncated URL. We do not expect to encounter any personally identifiable information in the truncated URLs we are collecting, but if we do, we will delete it promptly.

We will NOT know the identity of a resident browsing the Internet and will NOT be viewing the content of the web pages a resident visits.

What do I need to do if I don't want to participate in the study?

Any web traffic that is encrypted – i.e., traffic that is sent via https:// or through a VPN connection – cannot be captured or recorded as part of this study. So if you decide you don't want any of your web browsing information to be recorded as part of this study, you can encrypt your web traffic in one of two ways:

- ensure that the website you are using has a URL starting with https:// instead of http://
- use IU's Virtual Private Network (VPN) service to do all of your web browsing. For more information about how to use the VPN, see http://kb.iu.edu/data/ajrq.html, or call the UITS Support Center at x5-6789.

There is no penalty for not participating in this study. Since no personal information will be recorded, we won't be able to tell whether a given resident even participated or not.

When will the study take place?

We will record information between [dates]

Has this study been approved by Indiana University?

Yes. The IU Bloomington Human Subjects Committee and Residential Programs and Services have both approved this study.

If I have questions about the study, who should I call?

For more information on this study or its procedures, please contact the Primary Investigator, John Duncan, at johfdunc@indiana.edu. Please put "Browser Study" in your email's subject line. If you wish to speak with the research team by phone, you may contact us at [phone number removed]. If you have any other concerns about the study or its procedures, please contact the office for the Indiana University Bloomington Human Subjects Committee, Carmichael Center L03, 530 E. Kirkwood Ave., Bloomington, IN 47408, [phone number removed], or by email at iub_hsc@indiana.edu.

A.6 – Final Version – 02/19/2008

Notice: Your residence hall has been selected to participate in an IU research study.

We are giving you this notice of our research study so that you may choose whether or not to participate. See below for a description of the study and what you need to do if you do not wish to participate.

Who are we and what are we studying?

Our research team at the IU School of Informatics is working to better understand what makes people vulnerable to fraud and other scams when using the Internet, and what types of solutions may help protect them against such threats. As part of these efforts, we have designed a study to measure two things. First, do people who are similar geographically and demographically have similar web browsing habits? Second, how well do traditional methods used to protect privacy truly protect private information? Studying these things will allow us to better assess existing technologies for protecting privacy, and whether a group of people in a social network can protect themselves from certain types of harmful web sites by voluntarily sharing information.

What information are we collecting as part of this study?

We will be recording a truncated version of the URLs (web addresses) of the web pages that residents request as the requests are transmitted over the IU network from your dorm to the Internet. We have chosen to study web browsing from a single dorm because the residents are by definition geographically similar, and they also tend to be similar demographically.

What information are we NOT collecting as part of this study?

We will NOT record any of the following:

- names, demographic information, or other personally identifiable information.
- the full URL or any content of web pages any resident browses. We will only record a truncated URL. We do not expect to encounter any personally identifiable information in the truncated URLs we are collecting, but if we do, we will delete it promptly.

We will NOT know the identity of a resident browsing the Internet and will NOT be viewing the content of the web pages a resident visits.

What do I need to do if I don't want to participate in the study?

Any web traffic that is encrypted – i.e., traffic that is sent via https:// or through a VPN connection – cannot be captured or recorded as part of this study. So if you decide you don't want any of your web browsing information to be recorded as part of this study, you can encrypt your web traffic in one of two ways:

- ensure that the website you are using has a URL starting with https:// instead of http://
- use IU's Virtual Private Network (VPN) service to do all of your web browsing. For more information about how to use the VPN, see http://kb.iu.edu/data/ajrq.html, or call the UITS Support Center at x5-6789.

There is no penalty for not participating in this study. Since no personal information will be recorded, we won't be able to tell whether a given resident even participated or not.

When will the study take place?

We will record information between March 5, 2008 and May 3, 2008.

Has this study been approved by Indiana University?

Yes. The IU Bloomington Human Subjects Committee and Residential Programs and Services have both approved this study.

If I have questions about the study, who should I call?

For more information on this study or its procedures, please contact the Primary Investigator, John Duncan, at johfdunc@indiana.edu. [phone number removed]. If you have any concerns about the study or its procedures, please contact the office for the Indiana University Bloomington Human Subjects Committee, Carmichael Center L03, 530 E. Kirkwood Ave., Bloomington, IN 47408, [phone number removed], by email at iub_hsc@indiana.edu.

Appendix B - Experiment Addendum

Addendum to Human Subject Study # **07-12550**

John F. Duncan, IUB CSCI, PhD student, johfdunc@indiana.edu

Dr. L. Jean Camp, IUB Informatics, Associate Professor, ljcamp@indiana.edu

To whom it may concern,

In our original filing of study # 07-12550, we included the following line in answer to question A of the Summary Safeguard Statement:

"After the study, data will be retained only on a single CD stored in a locked room."

We are writing to request a modification to this. It has come to light, now that we have collected the data and wish to begin processing it, that several additional difficulties will be caused by our original statement.

A large amount of data has been collected, which will only fit on a single CD in a highly-compressed format unusable for data processing. Thus, in order to process the data, we would have to offload it from the CD to some local storage repository, such as a workstation in CS or Informatics. Should we offload it in this manner, issues arise such as secure data erasure. Additionally, were this CD to be damaged in any manner, the entire study might collapse for loss of the data. Thus, we propose the following addendum to the study, which should be taken to replace the original line:

"After the study, data will be retained in two places. First, as a backup, several redundant copies will be stored on CDs in a locked storage facility. Secondly, a suitable secure network storage point will be chosen (IU offers several such central repositories, and we plan to use the MDSS – Massive Data Storage Service) and a copy of the data in non-compressed form will be kept there, accessible over the network by the researchers conducting this study, in order to allow them to process the data. In this manner, the data will remain only on machines that fall under UITS' network security and data protection policies."

We write to ask that you approve this highly necessary addendum to our study with all due haste so that we may proceed with our data analysis.