It's Not Stealing If You Need It: A Panel on The Ethics of Performing Research Using Public Data of Illicit Origin

Serge Egelman[a], Joseph Bonneau[b], Sonia Chiasson[c], David Dittrich[d], and Stuart Schechter[e]

[a]University of California, Berkeley
[b]University of Cambridge
[c]Carleton University
[d]University of Washington
[e]Microsoft Research

1 Introduction

In a world where sensitive data can be published to a worldwide audience with the press of a button, researchers are increasingly making use of datasets that were publicized under questionable circumstances. In many cases, such research would otherwise not be possible. For instance, Weir et al. examined over thirty million user-generated passwords in order to observe the effects of entropy on password cracking [10]. All of the passwords in their dataset were obtained from various private databases that were breached by others and then subsequently posted to the Internet, the vast majority of which came from the RockYou breach [9]. Komanduri et al. used this same dataset to examine the effects of password creation policies on entropy [5]. Research on how users generate passwords is important, as passwords are the most common authentication mechanism. The resulting publications help system designers create password policies that balance both security and usability. Such data is only available as the result of an independent party's illegal actions. At the same time, the question exists of whether benefiting from this data makes a researcher a party to the underlying release, and whether the resulting research is ethical. This is a difficult question, especially when similar data could not otherwise be gathered: passwords generated solely for study lack ecological validity and "real" passwords are usually unobtainable due to obvious security concerns. Thus, if the researchers are not personally involved with the illegal acquisition of goods, does their use create an ethical dilemma?

Similarly, some researchers have gone beyond simply using data that others have published. In the course of gathering data, many have likely violated various terms of service—civil contracts. Amitay published an analysis of iPhone unlock PINs that were collected by his app, which mimics the iPhone unlock screen [1, 3]. He published a summary of this data with the goal of demonstrating that users choose predictable PINs (e.g., 1234) and hoped that this may prompt them to choose more secure ones. This resulted in Apple removing the app from the

app store, alleging a breach of their agreement. Bonneau et al. published a study on the use of so-called "secret questions" used for backup authentication with the goal of making these questions harder to compromise [2]. Part of this research involved compiling lists of common names by crawling Facebook. Others have performed similar research involving crawling various social networking sites [4, 7, 6]. All of these studies likely violated the sites' terms of service. This raises another ethical question about where the line should be drawn: are there fewer ethical issues involved with gathering data by violating terms of service (i.e., civil law) vs. violating criminal laws?

The use of data of questionable provenance in research is not just limited to passwords. Graphics researchers routinely use a test image featuring a female model known as "Lenna" [11]. The origin of this image was from a November 1972 issue of *Playboy* magazine. Despite being copyrighted, this image routinely appears in journal and conference publications. While Playboy, the copyright holder, has not taken action against any researchers to date, the ethics and legality of this practice—despite being widespread—are still questionable. When an ethical violation has become pervasive, does that lessen its magnitude? Is it no longer unethical if it becomes a social norm?

These examples illustrate how the desire to disseminate knowledge for the greater public good may involve actions that are ethically debatable. Indeed, we are organizing such a debate. Our panel will focus on discussion surrounding the ethics of using stolen data for research purposes. The panel will be moderated and will feature panelists representing the following viewpoints:

- Someone who has used stolen data to conduct research.
- Someone who does human subjects research outside the US.
- Someone who sits on an Institutional Review Board (IRB).
- Someone who is morally opposed to using stolen data in research.

2 Participants & Positions

2.1 Joseph Bonneau

I advocate that we can adapt ethics of "white-hat hacking" to the use of illicit data in research. The research community generally accepts papers which identify vulnerabilities in real software or websites, subject to a few basic principles. I propose that we work to adapt these into a set of ethics for using illicit data. First, we should develop a "do no harm" principle which can be realized by only using illicit data to advance scientific knowledge and not aid any parties in acting maliciously. In many cases there are technical ways to transform illicit data to prevent illicit use while still enabling research, such as stripping usernames out of a leaked password file. Second, we can require responsible disclosure, which is easy to adopt and often superfluous if companies already know that they have lost data. Third, external review of proposed studies, for example by an appropriate institutional ethics board, can help researchers in designing ethical studies. It is important to develop these principles as studies involving leaked data become more prominent, but I believe the scientific potential of illicit data sets is too large to ignore their use.

Joseph is a PhD student at the University of Cambridge. His forthcoming thesis will focus on the statistics of human chosen secret distributions such as passwords, PINs, and passphrases. This research has included many real-world datasets, both leaked and obtained with permission. Joseph's prior research has included side-channel cryptography, obfuscation, reverse engineering, and white-box cryptography. Prior to his PhD, Joseph worked at Cryptography Research, Inc. He holds MS and BS degrees from Stanford University.

2.2 Sonia Chiasson

I think that as a research community, we need to come up with clear guidelines and minimum ethical standards for what we will accept for publication in international venues. These standards should be upheld regardless of whether the researchers' IRBs (in some cases these are non-existent) or local/national laws are more permissive.

I am not entirely opposed to using publicly available stolen datasets, but the case must be made for no conceivable harm to the victims. Cases where the "greater good" is served at the expense of a relatively small number of victims should not be entertained.

The issue of consent is important here — if we were to conduct a study to collect this same data rather than using a stolen set, would we need informed consent from participants? Should we require researchers to put in a reasonable effort at obtaining consent after the fact (they probably have usernames/email addresses available), if they want to use stolen data? It may be a daunting task, but perhaps this is the most ethical way to deal with the issue.

Sonia Chiasson is an assistant professor in the School of Computer Science at Carleton University in Ottawa, Canada, where she holds the Canada Research

Chair in Human Oriented Computer Security. Her main research interests focus on the intersection between human-computer interaction and computer security. Current projects are on user authentication, usable security for mobile devices, and computer games for teaching about computer security. She leads the NSERC ISSNet project on Human Behaviour and Computer Security. Before moving to Ottawa, she was an instructor in the Department of Computer Science at the University of Saskatchewan and a member of the HCI Lab. She has been conducting empirical studies requiring approval from ethics review boards for over a decade.

2.3 David Dittrich

The Common Rule has many definitions and proscribes what research is or is not exempt from IRB review. It is unclear how any given IRB would determine which question is more important: that research is exempt from review because the stolen data is "public" (45 CFR 46.101(b)(4)), or that there is personally identifiable information in the stolen dataset that was obtained illegally under circumstances where those persons identified reasonably believed their data was not being recorded and would remain private (45 CFR 46.102(f)(2)). I believe it is more important for researchers to always be able to clearly and coherently explain their intent in performing research using stolen data, who the researcher is trying to serve, what measures the researcher is taking to balance benefit to society vs. risk to those identified in the data, and how those individuals identifiable in stolen data will feel about the fact that their stolen data was made public, how it was studied and what about it was published.

David has over 15 years of experience in computer security operations, computer forensics, network forensics, distributed intruder attack tools (also known as "botnets"), and the legal and ethical frameworks for responding to computer attacks. He has co-authored several papers, articles, and book chapters dealing with legal and ethical issues in computer security research and operations. David has served on the University of Washington's IRB Committee K for the past two years, where he provided data security expertise to his Committee and occasionally to PIs.

2.4 Stuart Schechter

Just as one cannot assume that an act that has not been deemed illegal is socially acceptable, one cannot assume that research that is not forbidden by the common rule, and allowed by IRBs, would be considered ethical by greater society. Alas, the ethical debate over the acceptable use of stolen data often ends with a declaration that once the data becomes public, the rules of the game make its use acceptable. Consider, for example, if attackers who had compromised and released email passwords had also harvested emails and posted them publicly. Researchers might be tempted to use the data to determine if certain traits revealed in the emails (e.g., erectile dysfunction) were correlated with other, possibly more

embarrassing, traits (e.g., affinity to the music of Barry Manilow). Even if individuals who had written the emails being studied were not identified by the researchers and came to no personal harm, these unwitting research participants might consider it unethical that their personal information be used by researchers without their consent. Such a study could not be ethically justified purely on the willingness of an IRB to approve it. Similarly, it is not sufficient to assume that lists of compromised passwords are fair game so long as criminals have already made the lists sufficiently public. They must imagine all reasons why the owners of this passwords might object to the use of these passwords and argue why they feel justified in going forward despite these objections. Researchers should not treat compliance with rules as a substitute for sufficient ethical consideration, as doing so may lead to these rules causing more harm to participants than protection.

Stuart is a man of few accomplishments and so, the reluctant reader should be pleased to learn, his biography is correspondingly short. Stuart researches computer security, human behavior, and occasionally missteps in such distant topics as computer architecture. Those who have worked with Stuart rave about his "tireless dedication to shooting down any idea that he cannot take credit for." Institutions that may or may not be re-evaluating their admissions or hiring policies in response to past associations with Stuart include The Ohio State University College of Engineering (B.S.), Harvard's School of Engineering and Applied Sciences (Ph.D.), MIT Lincoln Laboratory (his former employer), Microsoft Research (his current employer), and KAIST (to use a Facebookism, "It's complicated").

2.5 Serge Egelman

Serge Egelman, normally type cast as an instigator, will be in the role of moderator. Expect a lively panel.

Serge is a postdoctoral researcher at the University of California, Berkeley. His research focuses on usable security, with the specific aim of better understanding how people make decisions surrounding their privacy and security, and then creating improved interfaces that better align stated preferences with outcomes. This has included human subjects research on social networking privacy, access controls, authentication mechanisms, web browser security warnings, and privacy-enhancing technologies. He received his PhD from Carnegie Mellon University and prior to that was an undergraduate at the University of Virginia. He has also performed research at NIST, Brown University, Microsoft Research, and Xerox PARC.

3 Post-Panel Summaries

3.1 David Dittrich

This panel looked at the question of whether or not it is ethical to use stolen data, made available on public web sites without the consent of the owners

of that data or anyone potentially exposed within the data, in research. Just because it is hard to get access to data, does not mean it is okay to use any data a researcher can get their hands on. Nor does it mean a researcher can take short-cuts that may increase risk to individuals who are identifiable in data used in research (regardless of whether or not those identified are the direct subjects of research).

Implicit in the question of the ethics of using publicly available stolen data is a determination of whether such data fits the criteria of "research using publicly available data sets," as well as whether such a determination by itself is sufficient for research to need Institutional Review Board (IRB) review (even expedited review of minimal risk research). Just because data is found on a web page does not make it "public." Researchers have been heard to utter statements like, "I am using public data, which does not require IRB approval, so there is no need for me to even talk to my IRB." Such statements imply the researcher knows best and that no outside review of their actions are necessary. The argument that researchers are capable of deciding for themselves what is or is not subject to external review is belied by stories of failed self-regulation of research in books like, *The Immortal Life of Henrietta Lacks* [8]. This book is widely read and discussed in the IRB community for its telling of the personal story of a family that suffered multiple medical research abuses in the mid 1900s. Researchers cannot always be trusted to act appropriately in the face of potentially harmful research and self-interests, which is part of the reason why IRBs exist today.

Private data that was obtained through illicit means (e.g., data stolen in an intrusion incident) and put on a public web site is still private data. U.S. Federal Regulation 45 CFR 46.102(f)(2) defines "identifiable private information" as including:

> "Information about behavior that occurs in a context in which an individual can reasonably expect that no observation or recording is taking place, and information which has been provided for specific purposes by an individual and which the individual can reasonably expect will not be made public (for example, a medical record). Private information must be individually identifiable (i.e., the identity of the subject is or may readily be ascertained by the investigator or associated with the information) in order for obtaining the information to constitute research involving human subjects."

Therefore, some data made publicly available, such as the Statfor subscriber database stolen by LulzSec/Anonymous in December, 2011, would fit the definition of "identifiable private information" and would likely require IRB review of use in research, regardless of whether that data is available on a free and open public web site like Pastebin. Data sets, such as the RockYou password file, may also fit this definition.

To a large extent, IRBs at each institution in the United States function independently and have some leeway to interpret/apply the elements of the "Common Rule" as they see fit. Each federally funded research institution in

the United States operates under something known as their Federal Wide Assurance (FWA). The FWA is the institution's commitment to the Department of Health and Human Services (HHS) that it will comply with HHS rules for human subjects protection under 45 CFR 46. Some institutions may choose to require IRB review for all research at the institution, regardless of the funding source, while others may only require that federally funded research go before an IRB. The IRB committee is there to evaluate the risk to subjects from the research subjects' perspectives, in a way acting as their representative.

Those who own the data, and those who are identified within the data, may have an expectation of privacy in that data. When stolen data is made public, and a private individual decides to archive that data, they are likely operating outside the purview of an IRB and may be taking no consideration of the risks to identifiable individuals that an IRB would. The researcher wanting to use that data may, however, be operating within the IRB's purview and must conform with institutional requirements for IRB review of proposed research. A situation in which researchers bypass IRB review by asserting the "public data" exclusion may create an environment where individuals purposefully steal data in order to make it available to researchers, which violates both the spirit and letter of the law regarding human subjects protection via IRBs.

The identifiability of individuals within the data may be of greater importance in evaluating whether an IRB committee must approve research using public data than simply answering the question, "is the data available to anyone on the internet?" There may be instances when publicly available data may be partially de-identified, but can be combined by a researcher with other data sources, re-identifying individuals within the data. The act of re-identifying individuals and exposing them publicly can be harmful to those individuals. For this reason, many bio-repositories that make de-identified data available to researchers without necessitating IRB review, in order to safeguard the identifiability of subjects, will require the researcher to sign an agreement that includes a clause that prevents the researcher from taking steps to re-identify the individuals whose bio-samples are being studied. While it may show cleverness on the part of a researcher to identify an individual from de-identified or anonymized data, a researcher could be sanctioned by their IRB for doing so.

The University of Washington's Human Subjects Division publishes guidance/policy on use of public data sets.[1] UW's policy defines public data sets as being, "data files prepared by investigators or data suppliers with the intent of making them available for public use" and discusses usage restrictions, access agreements for restricted datasets, and data protection mechanisms that must be applied to ensure no unauthorized disclosure of individuals who are identifiable within data sets. They also define what "publicly available" and "de-identified" mean. A list of over two dozen data sets that have been evaluated by the UW IRB office are on a list of approved data sets that require no IRB review. Researchers who want to use other data sets that are not on the pre-approved list can nominate the data set for evaluation. If a funding agency does not require

[1] http://www.washington.edu/research/hsd/docs/1125

IRB review for publicly available data, researchers can provide documentation to that effect and the IRB will make a determination about whether any IRB review is required. For all other data, the IRB evaluates the proposed use of the data.

It is not a researcher's right to decide whether their research is exempt from IRB review, or whether data they wish to use does or does not conform with the definition of "public data" under the Common Rule. The researcher is obligated to confirm their interpretation with the IRB, who is the arbiter of how the Common Rule is interpreted as specified in their FWA. The researcher may risk sanction if they bypass or ignore the IRB's determination, which can vary by institution and by IRB. OHRP is relatively silent on the parameters of non-compliance. If an IRB determines a researcher acted unethically, or failed to submit research or data use to review when it should have been evaluated, the IRB may have the authority to do any/all of the following: (1) Halt current research and/or any further research; (2) Ask for publication of results to be halted, withdrawn, or modified to note researcher non-compliance; (3) Cite the researcher for serious non-compliance; (4) Require that all future research by that researcher be reviewed.

In other words, when it comes to performing research using stolen data, the catch phrase should be "researcher beware."

3.2 Stuart Schecter

Stuart argued that exemption four in the Common Rule, which states that all research using publicly available sources need not be reviewed by IRBs, gives researchers the freedom to perform studies that a great majority of the public might consider objectionable and unethical. He cautioned that researchers should not "turn off their ethics caps" and assume a study will be considered ethical simply because it qualifies for exemption from IRB reviews. To support this position, he provided three examples of research that qualifies for exemption four, but for which the social costs may outweigh the benefits.

In the first example, he explained example passwords from a compromised password data set may be traceable back to the accountholder even if no other information about the accountholder is present. It may be a password that contains data about the accountholder or even a password that appears random, but that contains a string that others may associate with that accountholder. For example, part of the password may be a password that the user shares with a significant other.

In the second example, Stuart described the implications if researchers were to come across a publicly-available repository of thousands of stolen medical records. He described how researchers might use these records to create a machine learning algorithm that could predict the likelihood that a patient suffered from a degenerative mental illness that would cause increasingly erratic behavior. The consequences of such research is those patients who this algorithm indicates are likely to be suffering from this illness—but were not yet diagnosed—would

have their potential condition revealed to anyone who cared to run the algorithm on the data set. Stuart provided an example of a hypothetical individual, diagnosed with this degenerative mental illness, having to live the remainder of his life with every friend and colleague concerned that his every behavior might be the result of a mental condition predicted by this algorithm.

In the third hypothetical example, Stuart described how researchers might abuse a publicly-available repository of stolen health records from minority groups (e.g., racial minorities or LGBT). He described how researchers at religiously-affiliated anti-homosexual universities might use the data to argue that homosexual youth were more likely to engage in a socially undesirable behavior (e.g., smoking) or how other researchers might use data on racial minorities to associate them with genetically undesirable traits.

Stuart argued that in many of these cases, the general public would find such research objectionable and question any system of ethical regulation that exempted it from review.

In these cases, Stuart proposed that the standard of ethical behavior should rely on whether researchers could reasonably anticipate that the great majority of those whose data had been stolen would consent to the research taking place, and that the social benefits outweigh the social costs. Stuart's position is thus that public data should only be exempt from ethics reviews if the data were made public with the consent of its subjects.

References

1. Amitay, D.: Most common iphone passcodes (June 13 2011), http://amitay.us/blog/files/most_common_iphone_passcodes.php
2. Bonneau, J., Just, M., Matthews, G.: What's in a Name: Evaluating Statistical Attacks Against Personal Knowledge Questions. Financial Crypto '10: The Fourteenth International Conference on Financial Cryptography and Data Security (2010)
3. Bonneau, J., Preibusch, S., Anderson, R.: A birthday present every eleven wallets? The security of customer-chosen banking PINs. In: FC '12: Proceedings of the the Sixteenth International Conference on Financial Cryptography (March 2012), http://www.cl.cam.ac.uk/~jcb82/doc/BPA12-FC-banking_pin_security.pdf
4. Gross, R., Acquisti, A.: Information revelation and privacy in online social networks. In: WPES '05: Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society. pp. 71–80. ACM, New York, NY, USA (2005)
5. Komanduri, S., Shay, R., Kelley, P.G., Mazurek, M.L., Bauer, L., Christin, N., Cranor, L.F., Egelman, S.: Of Passwords and People: Measuring the Effect of Password-Composition Policies. In: CHI '11: Proceeding of the 29th SIGCHI Conference on Human Factors in Computing Systems. ACM Press, New York, NY, USA (2011), to appear.
6. Korolova, A., Motwani, R., Nabar, S.U., Xu, Y.: Link privacy in social networks. In: Proceeding of the 17th ACM conference on Information and knowledge management. pp. 289–298. CIKM '08, ACM, New York, NY, USA (2008)
7. Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., Bhattacharjee, B.: Measurement and analysis of online social networks. In: Proceedings of the 7th ACM

SIGCOMM conference on Internet measurement. pp. 29–42. IMC '07, ACM, New York, NY, USA (2007)
8. Skloot, R.: The Immortal Life of Henrietta Lacks. Broadway (2010)
9. Vance, A.: If your password is 123456, just make it hackme. New York Times, http://www.nytimes.com/2010/01/21/technology/21password.html (January 2010, retrieved September 2010)
10. Weir, M., Aggarwal, S., Collins, M., Stern, H.: Testing metrics for password creation policies by attacking large sets of revealed passwords. In: Proceedings of the 17th ACM conference on Computer and communications security. pp. 162–175. CCS '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1866307.1866327
11. Wikipedia: Lenna (Accessed: November 2, 2011), http://en.wikipedia.org/wiki/Lenna