;login:
DECEMBER 2013 VOL. 38, NO. 6
ELECTRONIC SUPPLEMENT

Conference Reports

CSET '13: 6th Workshop on Cyber Security Experimentation and Test

Washington, D.C.
August 12, 2013

Panel

Conducting Research Using Data of Questionable Provenance

Summarized by Hadi Asghari (hasghari@acm.org)

Panelists: Michael Bailey, University of Michigan; Lujo Bauer, Carnegie Mellon University; L. Jean Camp, Indiana University; Sven Dietrich, Stevens Institute of Technology; Damon McCoy, George Mason University.

Moderator: Chris Kanich, University of Illinois at Chicago

This session was an exciting and interactive discussion where the panelists and audience exchanged views on how to deal with data sets that are "questionable" from an ethical point of view. Questionable data sets are those whose origins are unknown, or are morally ambiguous, such as leaked data sets and data gathered without consent. The session started from the ethical implications of using such data, and moved on to how the community could reach a consensus on what the ethical norms are and they should be policed.

The panel began with Chris Kanich asking all panelists for opening comments. Michael Bailey and Damon McCoy emphasized that this is a very important topic given current events, emerging trends, and more and more data. The community faces challenges and needs to come up with guidelines to defend itself from attacks by outsiders on its ethical stances. Lujo Bauer gave the example of leaked password sets that he used in his work. He said that he was initially puzzled by reviewers asking him about the ethical implications of using data obtained "via theft" for research purposes, but now finds this a valid question. Sven Dietrich had similar experiences of using "live" data in their work on botnets. Jean Camp highlighted that more established scientific fields have clear ethical guidelines for research, and gave the example of a special issue of Nature devoted to the matter, while computer security research does not. In her opinion, the ethics of computer security research differ from other fields "because HIV does not become more potent as a result of disclosure, whereas our intelligent adversaries do." Lujo Bauer didn't fully agree with this point.

The floor was next opened for the audience. Stefan Savage got the ball rolling by asking the panelists whether it is ethical to use a leaked data set such as WikiLeaks. Michael Bailey's response was that this depends on the data set and on how the benefits and harms balance each other. He said one relevant question might be, is it a unique set, or one that could have been obtained by less questionable means? Referring to the Carna botnet, he said it's just as important to know what benefits will come out of the data as to understand the harms but that one would need to consider whether sensitive data—people's terminal illnesses or sexual preferences, for example—had been revealed. Damon McCoy had a similar opinion and believed that we need to look at the affected stakeholders and how they could be protected against potential harms; and at the benefits of including the data in a report. He also pointed out that one of our duties as researchers is to validate the authenticity and correctness of a data set we want to use; for leaked data sets, this is not easy, as the methodology is often not documented, and nobody's reputation is on the line.

Lujo Bauer agreed that the scientific quality of leaked data is not always good, given his own experience with the leaked passwords, partly due to the nontransparent methodology. He also emphasized looking at the benefits to society as a whole when doing cost-benefit analysis, not just to the benefits for the individual researcher. A member of the audience (one of the authors of the ZMap tool) voiced concern that using leaked data sets might encourage unethical data gathering practices in the future. Jean Camp, highlighting the Carna botnet data, stated that she would have difficulty accepting any paper based on it. She believed that the data was not unique, the Census paper did not convey professionalism or reflection, and there were obvious ethical and legal implications to the data-gathering process, so overall, using it was ill-advised. Sven Dietrich backed this opinion and asked whether we can actually distance ourselves from the "crime" committed for collecting the data and just reap the benefits of using it? Michael Bailey didn't agree and countered that since the harm had already been done and the data set was public, we could use it. Some members of the audience questioned whether all the harm was really done. Michael agreed that perhaps breach of trust and damage to community reputation might be a new harm.

Jean Camp then raised her own question: how can we as a community move forward on this issue, putting aside individual values, and say this data or practice is obviously fine, this needs to be motivated to show that it has a genuine contribution to the science of cybersecurity, and this is beyond acceptance. Otherwise, the community might face backlash from society or government, with reactions such as the Computer Abuse and Fraud Act. Stefan Savage consolidated one side of the panelists' view as follows: the strict utilitarians on the one side, which is also the more common stance. Stefan Savage then asked in what units do we measure utility. On the other side were the more deontologically focused people who say it is just bad practice, end of discussion. Since these views are hard to reconcile, different communities will end up adopting different lines, and people in the communities will end up voting with how they review papers.

This started the second discussion topic: the role of program committees (PCs) in enforcing ethical rules. Lujo Bauer asked the panelists and audience if they thought program committees could reject a paper on ethical grounds. He mentioned being caught in several such discussions in PCs recently. His personal opinion was that this should be part of the committee's job and chairs should announce this publicly. Several people from the audience reacted. Zachary Tudor asked whose ethical rules should be enforced; another audience member thought this will result in the community having norms that differ from what they think is acceptable. Stefan Savage thought this would actually be perfect, as we get to see a unique signal: at each point in time, what papers were deemed fine by each PC. Someone asserted that PCs should in such cases inform authors so that they can learn from it. Peter Johnson followed up on this by asking if universities should be notified when a paper gets rejected for ethical reasons.

This created some discussion among the panelists, but the general consensus was that no, unless some additional harm might get done, for instance if IP addresses of individuals might be published. Jean Camp gave the example of what they were doing and worked well in the Symposium on Usable Privacy and Security (SOUPS) community: ethical principles are explicitly mentioned in the call for papers. Moreover, she hoped that at some point in the future, the US ACM or a similar organization would provide clear guidelines on ethical issues. This way, researchers would be able to know before submission to what ethical standards they would be held to, instead of having to learn ethics via gossip.

The conversation then shifted to a third topic: the role of university institutional review boards (IRBs). José Fernandez raised the question to the panel of whether they thought computer security research should be part of RB mandates, as they currently are not in the majority of places, despite being a risk. Sven Dietrich responded first. They had added it to their university's IRB mandate, but when the board later changed, it was again removed. Lujo Bauer pointed out that although educating IRBs is a good idea, it does not solve the main issue of whether questionable data sets can be used. Michael Bailey's view was that IRBs won't be helpful on this matter, and will probably just say it's public data, so it's OK. Stefan Savage pointed out that much corporate research (e.g., Microsoft or Google Research) and many institutions don't even have IRBs; additionally, IRBs are not standards bodies, but check conformance. Jean Camp pointed out that IRBs can help those who want to be ethical, but in the end we need to police ourselves.

The discussion then flowed back to the issue of program committee roles. A member of the audience asked whether ethic violations should be publicized by PCs? Stefan Savage questioned how that might be even possible, as publishing the papers would be unacceptable. Michael Bailey was also negative, arguing that we don't publicly talk about other problems in a paper, for instance when a data set is not large enough, so why should we do it here? One audience member responded that ethical violations cause harm, but statistical errors don't. The discussion continued for a while. Jean Camp was in favor of increasing transparency without shaming—for instance, by presenting statistics of rejections or giving generic statements on previous rejected papers (e.g., we reject papers with the word "X." She also mentioned a norm instantiated at SOUPS: every paper has to have a section on the ethical implications of the research and data. This requirement was added because of cases where papers got rejected on ethical grounds when in fact the authors had already gotten the backing of IRBs (or the equivalent) for the research, but failed to discuss this. Sven Dietrich and Michael Bailey agreed that requiring ethics to be directly addressed by authors could be a good way forward for chairs of other program committees. Michael added that, in his opinion, the ethics section should be part of the methodology section of all papers, as "it is part of the science." Jean Camp and Lujo Bauer strongly approved.

After one hour and twenty minutes of fascinating debate, Chris Kanich asked the panelists for their concluding and closing statements. Sven Dietrich believed that raising awareness in the community was the way to go, be it via IRBs, ethic committees, program committees, or otherwise. Jean Camp strongly advocated openness and documentation, and was against shaming. Giving clear feedback as soon as possible to grad students on the ethical aspects of research is also important to avoid "nightmare scenarios." Lujo Bauer said he particularly liked the "ethics is part of the science" statement. He also said that the issue should be dealt with by PCs and not passed along to others, such as IRBs. Damon McCoy concluded that addressing the ethical issues and arguments of a research in the methodology will help others read and learn from it, especially since there seems to be no one solution at this point. Michael Bailey said that he found the discussion to be very fruitful. Chris closed the session by thanking the panelists and audience for their participation.