Ethical issues in research using datasets of illicit origin

Daniel R. Thomas       Sergio Pastrana       Alice Hutchings
Richard Clayton       Alastair R. Beresford

Cambridge Cybercrime Centre, Computer Laboratory, University of Cambridge, UK

Internet Measurement Conference (IMC) 2017

GPG: 5017 A1EC 0B29 08E3 CF64 7CCD 5514 35D5 D749 33D9
Think before using data of illicit origin

- Include an ethics section
- “human participants” not “human subjects”
- Talk to your Research Ethics Board (REB) (IRB / Ethics committee)
We had questions about ethics in our research\(^1\)

- UDP honeypot DDoS sensors
- Developed statistical method for estimating coverage
- Verified using leaked booter databases...

Case studies

- Malware & exploitation (Carna, ...)
- Password dumps
- Leaked databases (Booters, Patreon, ...)
- Classified materials (Snowden, Manning)
- Financial data (Panama papers)

We drew out common justifications, safeguards, potential harms, and benefits.
Ethics are norms of conduct

- Distinguish between acceptable and unacceptable behaviour
- Guidance in Menlo Report\textsuperscript{2}
- Enforced by REBs and Program Committees (e.g. IMC\textsuperscript{3})
- Minimise harm, maximise benefit.


\textsuperscript{3}Mark Allman and Vern Paxson. 2007. Issues and etiquette concerning use of shared measurement data. In Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. ACM, 135–140.
We should consider these ethical issues

- Identification of stakeholders
- Informed consent (where possible)
- Identify harms
- Safeguards
- Justice
- Public interest
## Leaked databases: Ethical issues

<table>
<thead>
<tr>
<th>Sources</th>
<th>Ref</th>
<th>Year</th>
<th>Identify stakeholders</th>
<th>Identify harms</th>
<th>Safeguards</th>
<th>Justice</th>
<th>Public interest</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 carding forums</td>
<td>[12]</td>
<td>2013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TwBooter, 14 others</td>
<td>[9]</td>
<td>2013</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Vdos, CMDBooter</td>
<td>[10]</td>
<td>2017</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>4 underground forums</td>
<td>[8]</td>
<td>2017</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Summary of the ethical issues and the identified by the authors for each paper. ✓ means the issue was discussed, ✗ means it was not.
Legal issues may arise

- Computer misuse
- Copyright
- Data privacy (e.g. GDPR)
- Terrorism
- Indecent images
- National security
- Contracts

IANAL. Just pointers: not even the beginning of advice.
AT&T iPad users database breach

- “Researchers” from Goatse Security found vulnerability in AT&T website
- Obtained email addresses for 114,000 iPad users
- Shared vulnerability, exploit, and email addresses with third parties
- Did not report vulnerability to AT&T
- Went to jail

Unethical and illegal.
Illegal but ethical research?

- **Mens rea**: lack of criminal intent
- Not in the public interest to prosecute
- REB approval! REB protection?
- Get the law fixed
Bad and good justifications

- Not the first
- Public data
- No additional harm
- Fight malicious use
- Necessary data
Leaked databases: Justifications and oversight

<table>
<thead>
<tr>
<th>Sources</th>
<th>Ref</th>
<th>Year</th>
<th>Not the first data</th>
<th>Public data</th>
<th>Fight malicious use</th>
<th>Necessary data</th>
<th>Ethics section</th>
<th>REB approval</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 carding forums</td>
<td>[12]</td>
<td>2013</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>TwBooster, 14 others</td>
<td>[9]</td>
<td>2013</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Patreon</td>
<td>[7]</td>
<td>2016</td>
<td>X</td>
<td>✓</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
<td>Ø</td>
</tr>
<tr>
<td>Vdos, CMDBooster</td>
<td>[10]</td>
<td>2017</td>
<td>X</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>E</td>
</tr>
<tr>
<td>4 underground forums</td>
<td>[8]</td>
<td>2017</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

● means that the authors decided that the work could not be justified. Ø means not applicable and E means exempt.
Patreon database leak

- Researching crowdfunding on Patreon by scraping the website (complete scrape?)
- Patreon is compromised and full database + source code leaked
- Authors decided there would be additional harm
  - Private vs. public data
  - Legitimize criminal activity
  - Violate privacy

---

Safeguards against potential harm

- SS  Secure storage
- P  Privacy
- CS  Controlled sharing
Potential harms

- Illicit measurement
- Potential Abuse (PA)
- DeAnonymisation (DA)
- Sensitive Information (SI)
- Behavioural Change (BC)
- Researcher Harm (RH)
Researchers are participants

Ethics also protects researchers

Maltese journalist Daphne Caruana Galizia killed by car bomb used Panama Papers data also used by researchers
Benefits provided

- R Reproducibility
- U Uniqueness
- DM Defence Mechanisms
- AT Anthropology & Transparency
Leaked databases: Safeguards, harms and benefits

<table>
<thead>
<tr>
<th>Sources</th>
<th>Ref</th>
<th>Year</th>
<th>Safeguards</th>
<th>Harms</th>
<th>Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 carding forums</td>
<td>[12]</td>
<td>2013</td>
<td></td>
<td></td>
<td>DM,AT</td>
</tr>
<tr>
<td>TwBooster, 14 others</td>
<td>[9]</td>
<td>2013</td>
<td>P</td>
<td>SI</td>
<td></td>
</tr>
</tbody>
</table>

Safeguards: SS Secure Storage, P Privacy, CS Controlled Sharing.
Safeguards, harms, and benefits

Privacy: No individuals identified
Controlled Sharing: Under legal agreement
Sensitive Information: Names, email addresses, criminal activity
Behavioural Change: Don’t advertise (working) booters
Uniqueness: Ground truth on booter activity
Anthropology & Transparency: Real behaviour of booters

---

Let's do ethical research

- Explain why our work is ethical
- Learn from each other
- Improve Research Ethics Board (REB) process
Thank you! Questions?

Daniel.Thomas@cl.cam.ac.uk
https://www.cl.cam.ac.uk/~drt24/
GPG: 5017 A1EC 0B29 08E3 CF64 7CCD 5514 35D5 D749 33D9

Cambridge Cybercrime Centre has data to share
https://www.cambridgecybercrime.uk/

University support for complex ethics workshop
Oxford 12th December

Daniel R. Thomas was supported by a grant from ThreatSTOP. All authors are supported by the EPSRC [grant number EP/M020320/1].


