ETHICS GUIDANCE

TOWARD ETHICAL RESEARCH USING DATASETS OF ILLICIT ORIGIN

This document is a summary of:

Thomas, D. R., Pastrana, S., Hutchings, A., Clayton, R., Beresford, A. R., (2017) Ethical issues in research using datasets of illicit origin, *Proceedings of the 2017 Internet Measurement Conference,* London, United Kingdom, pp. 445-462, ACM. http://delivery.acm.org/10.1145/3140000/3131389/p445-thomas.pdf?ip=134.220.202.120&id=3131389&acc=ACTIVE%20SERVICE&key=BF07A2EE685417C5%2E300EF34F9F006C06%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&__acm__=1546877604_a09903332e135ee9b50692342dc42a82

and includes some information from

Poor, N. and Davidson, R. (2016) The ethics of using hacked data: Patreon's data hack and academic data standards. Data and Society. Council for big data, ethics and society, Case Study 03.17.16, March 2016), 1–7. https://bdes.datasociety.net/wp-content/uploads/2016/10/Patreon-Case-Study.pdf

An infographic providing details of quite a few examples of leaked data is available at http://www.informationisbeautiful.nt/visualeizations/worlds-biggest-data-breaches-hacks/.

1) Background

We may need to consider requests for ethical approval from researchers using data that was obtained without the consent of the original data owners or data subjects. There are three main issues concerning the use of such data:

1. Although data subjects gave consent for the original collection and use of the data, they have not given consent for its use in subsequent research.
2. Key stakeholders (the "leakers" or hackers) often have not sufficiently protected the identity/confidentiality of data subjects.
3. Key stakeholders may have broken the law in acquiring and distributing the data.

Despite these issues, it may be possible to use this type of data in research if additional safeguards are implemented to minimise potential harms, to protect data subjects and other stakeholders (*stakeholders* defined in Section 4) and, where possible, acquire informed consent from primary and secondary stakeholders.

The scientific method relies on empirical evidence to test hypotheses. Gathering and use of data is essential and supports evidence-based decision making. Researchers make significant use of data to support research and inform policy. This may include data obtained through illegal or unethical behaviour.

For Thomas et al. (2017) *datasets of illicit origin* are datasets collected as a result of:

- Exploiting a vulnerability in a computer system,
- Unintended disclosure by the data owner,
- Unauthorised leak by someone with access to the data.

Collection, use, or reuse of such datasets by researchers can have some advantages:

- Legitimate access to the data may not be possible,
- Use of datasets of illicit origin requires fewer resources than collection of data from scratch,
- The fact that this data can be shared and reused aids reproducibility.

There are many different types of datasets of illicit origin. We may need to take a different approach to each of them.

2) Topics of ethical concern
Thomas et al (2017) identify 2 topics of concern in the use of datasets of illicit origin.

- **Informed consent:** Participants' rights to withdraw and right to anonymity cannot easily be upheld when datasets of illicit origin are being used. These datasets may be large and have high dimensionality, making it easy to identify data subjects, even when anonymisation has been attempted. If consent was given on the basis of a promise of confidentiality, subsequent researchers should strive to maintain this.
- **Human rights** (to life; to freedom from arbitrary arrest; to a fair trial; to presumption of innocence before proof of guilt; to not having arbitrary invasions of privacy; to not being arbitrarily deprived of property): These may be compromised for data subjects in datasets of illicit origin because, as noted earlier, the data is large enough that data subjects can be identified. For example:
  - Identification of data subjects who use narcotic drugs in the Philippines creates the risk of their being targeted for extrajudicial assassination.
  - The leaking of 300,000 emails from Turkish Prime Minister Recep Erdogan exposed the home addresses of almost all adult women in Turkey. These details were also exposed about members of the ruling AKP party who could become future political targets (following the recent coup).

3) Guidance Checklist
Thomas et al. (2017) provide a guidance checklist relating to the release and use of datasets of illicit origin. They propose:

1. Research using such data must have a clear benefit to society. Ethical approval should be sought because it may contain identifiable information and the interests of data subjects must be protected.
2. The purpose and scope for using such data must be stated explicitly.
3. Privacy protection is a key ethical consideration
   - Data of illicit origin is unlikely to be anonymised.
     - Anonymisation of data is practically impossible (difficult, potentially expensive, and may change the data in important ways, making it somehow "artificial"), especially when it has high dimensionality and the dataset is large:

       "*The data science version of 'I'll go check the basement' in horror movies is 'Oh, it's OK, we anonymised the data.'*"
       -twitter user @lyda, following the release by Strava of data exposing the locations of secret military bases and patrol patterns in Syria

   - Raw data should not be shared publicly.
   - Research using this data should aim to preserve privacy.
   - Researchers holding datasets of illicit origin should not share the data:
     - Share only details of the source of the data.
       - If this is not possible then only share the data under a written *acceptable use policy*.
4. Research involving hacking into computers (including use of botnets) is usually unethical and illegal (computer misuse). Some justification may be offered (e.g. to conduct research in the hacking and "taking down" of websites used by criminals to support phishing) but it should not usually be accepted.

4) Questions to be Answered when Research Uses Datasets of Illicit Origin
Thomas et al. (2017) define the following types of stakeholders, who should be considered if research using these datasets is to be ethical:

- **Primary** – data subjects, people identified in the data,
- **Secondary** – intermediaries in the delivery of benefits or harms (e.g. service providers),
- **Key** – the leaker or researcher critical to the conduct of the research.

Then the following questions should be answered:

1. Is informed consent being obtained from primary and secondary stakeholders? If this is not possible, can the research be designed in such a way that it is not necessary?
2. What are the potential harms of conducting the research and using the data of illicit origin?

3. Which policies or mechanisms are in place to minimise the potential for harm or mitigate harm?
4. Does the research unfairly advantage or disadvantage any particular social or cultural group?
5. Is the research in the public interest? Is it publicly acceptable?

5) Legal Considerations
Does the research (data collection) involve an illegal activity such as hacking?

Does the research involve breach of copyright?

Data privacy – In some countries, personally identifiable information must be processed and protected according to relevant data privacy and data protection rules, e.g. GDPR:

- Personal data should not be included in publications.
- Data can be used for historical and scientific research regardless of the original purpose for which it was collected.
  - This seems to imply that, according to GDPR, subsequent use of the data for this purpose does not require informed consent.
- Data subjects must be protected.
- Information about data collection, safeguarding, and details of the way it will be processed must be made publicly available.
- Safeguards must be applied:
  - Encryption,
  - Pseudonymisation,
  - Data minimisation – redaction of information in the dataset that is irrelevant to the research question. This might be expensive in large datasets.

Does the research involve terrorist material (e.g. planning information)?

- In the UK, it is an offence to fail to report terrorist activity, including any discovered in a research project.

Does the data of illicit origin (which may include data scraped by robots from websites) contain indecent images of children?

Is the data classified? Even publicly accessible data may be classified. Extracts of such data disseminated at conferences can pose difficulties - In one case a video recording of a presentation at a conference was required by the US government to be destroyed because it included extracts of classified material.

Are there any legal contracts in place preventing the use of the dataset in this research?

6) Examples of research using datasets of illicit origin
I recommend reading more about these examples in Section 4 of Thomas et al. (2017). My summaries omit a lot of information.

6.1 Research Using Malware and Exploitation (of vulnerabilities in software)
In this first case, the data consists of computer programs:

- with vulnerabilities that may be exploited
- that may be illegal but are made freely available for study or misuse by others

Name: Carna scan
Purpose: To conduct a census of internet web pages
Methods: Used a botnet (processors in other people's computers being used *en masse* without their permission)
Ethical?: No. This was an example of *computer misuse.* The research made no technical contribution due to methodological errors made by the researchers.

Name: AT&T database of iPad users' email addresses
Purpose: To build a database of iPad users' email addresses

Methods: Researchers used a web service unwittingly made available by AT&T to obtain 114,000 email addresses of iPad users. They passed this information on to third parties without informing AT&T
Ethical?: No. One author was sentenced to 41 months imprisonment. Any benefit brought about by exposing the vulnerable web service was offset by the fact that the owner of the web service was not informed. This research was not in the public interest because harm was not minimised and benefit was not maximised.

Name: Use of leaked malware source code
Purpose: Aid understanding of how malware works to facilitate defense.
Ethical?: The purpose is in the public interest. However, possession of malware source code is illegal and accidental disclosure of it would be a harm. This research can be ethical if safeguards are in place:

- secure storage of the code
- enforced retention policies
- prohibition on public distribution of the source code
    - Authors can be identified from their code, making it possible to attribute malware attacks

6.2 Research Using Password Dumps
This category consists of huge lists of usernames, websites, and the associated password for each username. [See Thomas et al. (2017), Section 4.2]

6.3 Research Using Leaked Databases
Name: Booter databases (which contain distributed denial of service providers which can be used to attack websites and put them out of action). An example is bootyou.net. The user enters a web address, pays a fee, and the service attacks it until it goes offline.
Purpose: To better understand this criminal ecosystem
Methods: Study of a leaked database holding information on the activities of users of the booter services.
Ethical?: Yes, despite the fact that these databases are almost always illegal. The researchers used safeguards to avoid publishing personally identifiable data. They used publicly leaked data. No other ground truth on DDOS attacks is available, making the data unique. They identified potential harms and used safeguards.

Name: Patreon crowd-funding database
Purpose: To investigate the influence of users' social networks on their success in fundraising.
Methods: Patreon's database, containing data on projects, private messages, source code, email addresses, and passwords, was hacked and made publicly available.
Ethical?: Researchers declined to use this dataset, claiming it would be unethical:

- Risk of viewing private data unintentionally,
- Risk of legitimising criminal activity,
- Risk of violating users' expectations of privacy,
- Use of the data would be without the consent of the data subjects,
- Use of the data is unnecessary – some scraping of webpages from the Patreon website was deemed sufficient for researchers' purposes.

Name: Underground forums
Purpose: Underground forums are used by people to discuss illegal and criminal topics (among others). This leaked database included messages, usernames, user passwords, user credit card details, and more. The data can be used to gain insights into how markets for stolen data work and insights into other topics of discussion, many of which are criminal.
Methods: A forum of this type was hacked and made available.
Ethical?: In Thomas et al.'s (2017) review, none of the surveyed research conducted in this area was ethical. Although the purpose is in the public interest, none of the research exploiting this data mentioned the use of safeguards to protect the data, which was acquired illegally (through hacking). There is a risk of harm to identifiable data subjects (prosecution and physical threat). Due to the nature of the forum, it is likely to be impossible to acquire informed consent from the data subjects.

Type: Financial data leaks detailing financial behaviour of companies and individuals and peering arrangements between internet service providers

Name: Panama/Mossack Fonseca papers leak – internal database of the Panamanian law firm Mossack Fonseca leaked first to a German newspaper (Suddeutsche Zeitung) and then to the International Consortium of Investigative Journalists (ICIJ)
Purpose: To obtain information on criminal and unethical activities by numerous individuals and companies
Ethical?: The purpose is in the public interest. Research using this data can help make money laundering and tax evasion more difficult. However, not all of Mossack Fonseca's clients were engaged in criminal or unethical activities - there have been reports of undesirable unintended consequences for some of these "innocent" clients. Mossack Fonseca's work involves selling financial products to avoid taxes, which is unethical. Members of the ICIJ used safeguards to protect the data and the investigation. Ethical considerations concerning the use of this data have tended not to be discussed by researchers but arguments that the work is in the public interest currently prevail. Some uses of the data may be unethical in the same way that some uses of Mossack Fonseca's services may be unethical.

6.4 Research Using Leaked Databases (Classified Materials)

Type: Research using datasets of illicit origin which include classified materials detailing war decisions, espionage, and diplomatic activities
Name: Manning's WikiLeaks Dump – 700,000 documents and diplomatic cables from US government systems leaked to WikiLeaks. Full and un-redacted cables were released.
Purpose: To gain insights into the activities discussed in these diplomatic cables.
Ethical?: Use of this dataset in research is controversial, with no consensus in the academy on the morality of using the information – whether Manning is a traitor or freedom fighter, whether use of the data is ethical. Most researchers have so far preferred not to address this question.

Name: Snowden's NSA data leak - Large amounts of data from NSA and GCHQ leaked to journalists.
Purpose: To highlight pervasive surveillance of the public by the NSA.
Ethical?: Privacy and security of identifiable data subjects was not protected. However, the leak revealed pervasive surveillance by NSA, including the fact that the NSA had access to private data in cloud servers. This included data about clients of members of the American Psychological Association. Many view Snowden's actions as ethical civil disobedience. However, NSA claim that they were not involved in pervasive surveillance because the data was only processed by computer, not read by humans. The NSA claim that journalists acted unethically by reading this data themselves. Much of the data, although publicly accessible, remains classified. On occasion, agents of the US government have required that videos of academic conferences presenting extracts of classified data be destroyed.

7) Common Justifications from Researchers Using Datasets of Illicit Origin

These should be assessed objectively:

- "*We're not the first to use this data.*"
  - Was prior research ethical?
- "*This data was publicly available.*"
  - Is anonymisation needed?
  - Is it classified?
  - Poor and Davidson (2016) noted that when depositing the data, Patreon users had an expectation of privacy but following the Patreon hack, this expectation could no longer hold. They asked whether researchers need to respect the intent or the reality.
- "*Use of this data causes no additional harm.*" (Harm has already been done by the free distribution of the data. Our research doesn't directly add to this).
  - Will the research identify natural persons?
  - Will the data be kept securely?
  - Do the likely benefits of using the data outweigh the risks of causing additional harm?
  - Does the data contain obscene images that will be viewed by researchers?
    - In cases of child abuse, each viewing is considered an additional abuse
- "*Use of this data helps to fight against malicious use of it*"
  - Will the research use the data to combat use of the data by malicious parties?
- "*The information contained in this data isn't available anywhere else.*"

Thomas et al. (2017) note that one risk in much current (2017) research is that research using datasets of illicit origin passes ethical review because the research does not involve human subjects, just a dataset. As noted earlier, when datasets of illicit origin are used in research, it is important to consider its stakeholders.

8) Actions for research proposals using datasets of illicit origin in an ethical way

- **A1. Describe safeguards to avoid further disclosure of sensitive information**
  - Will the data be stored securely? Which type of encryption?
  - Will the research reveal identities of data subjects? Which method of anonymisation will be used, if any?
  - If data is to be published, which legal agreements will be used to prevent harms?
    - is the data being anonymised
    - is it being published only partially?
  - Will the research be published?

- **A2. Describe potential harms that may arise as a result of this research**
  - Does the research involve illicit measurement via hacking or paying offenders?
  - Could research results from this data be used by malicious actors to cause additional harm? (e.g. insights enabling improvement of malware or password hacking methods)
  - Could research using the data de-anonymise it? Are exposed data subjects at risk of prosecution or physical threat?
  - Does the dataset contain sensitive data such as passwords or home addresses that could be used to harm natural persons?
    - The size of the datasets can make this question difficult to answer. To illustrate, in plain text, 1 GB of plain text would be roughly one million pages of A4 text (500 million words at ~500 words per page).
  - Could exposure to the dataset cause harm to the researcher – prosecution, physical threat from criminals, emotional trauma, etc?
  - Might the research encourage future collection or use of data of illicit origin?

- **A3. Describe benefits of using the dataset**
  - Does use of the data make the research more easily reproducible?
    - Allows comparison of different algorithms or tools (in computational analysis)
    - For this to occur, *controlled sharing* may be required for data containing sensitive information
  - Is the data unique?
    - This is only a benefit is the data is actually useful!
  - Does the research facilitate development of defence mechanisms? (e.g. through study of new forms of cybercrime or the underground economy)
    - This may allow defences to be developed such as better anti-malware tools and better password policies
  - Does the data contain ground truth on human behaviour that would otherwise only be accessible in a filtered/biased form?
    - e.g. human behaviour when creating passwords
  - Does the data provide transparency through information that aids understanding of
    - Government surveillance methods
    - Company behaviour

    This may provide greater benefits than information about individuals alone. It may have additional public benefit by providing checks and balances on power

- **A4. Include an Ethics Section in your Paper**
  - Thomas et al. (2017) take the view that papers using data of illicit origin should always have an ethics section to explain how data was obtained, how it has been protected, to analyse the potential harms, benefits, and need for using such data. If data is being used that was shared under an acceptable usage policy, they propose that researchers should cite the policy that they are operating under.