Working with Sensitive Data in the Wild

Richard Clayton
richard.clayton@cl.cam.ac.uk

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

30th November 2015

What sort of data am I talking about ?

- I am going to be talking about cybercrime research because that is (mainly) what I do.
  - I also have lots of cybercrime data so I may be able to help you do cybercrime research as well
- But lots of other types of data is sensitive
  - medical devices
  - browser logs
  - smartphone usage patterns
  - commercial secrets
- Lots of other types of data is ‘wild’
  - measuring network properties
  - scanning the Internet for X
  - working out the location of the Uni 4 bus

Ethical approval

- You will almost certainly need ethical approval for research on wild or sensitive data. If you don’t obtain ethical approval then you may inadvertently do bad things
    - and THE PAPERS YOU WRITE MAY NOT BE ACCEPTED
- You also need to make peace with yourself that you are happy tackling a research topic rather than doing something else
- If you work on cybercrime you also need to consider the personal risks of looking closely at what the bad guys are doing
    - don’t want to overstress this, but some in this space do worry
- So before you can even start to fill in the ethics form you need to think hard about what you will be doing and how you (and others) are going to stay safe and happy

Next section of this talk is based on a peer-reviewed paper from an Ethics workshop:

Tyler Moore & Richard Clayton:
Ethical Dilemmas in Take-down Research
Second Workshop on Ethics in Computer Security Research (WECSR 2011), St Lucia, 4 March 2011

Phishing research

- Phishing is the theft of credentials by the use of fake websites
  - though ‘phish by email reply’ is also relevant these days
- We have a series of papers measuring this from 2007 onwards
  - all are branded “T. Moore & R. Clayton”
- So we are studying the actions of criminals – but we mainly studied (and carefully measured) the main countermeasure which is the ‘take-down’ (the removal of) the fake websites
- This ‘take-down’ was first done by the banks themselves, but various ‘brand-protection’ companies now do most of the work
- The WECSR paper sets out nine ethical issues that we have run up against during the course of our research. It’s ‘war stories’ rather than philosophy!

Dilemma 1: Should researchers notify affected parties in order to expedite take-down?

- We were measuring take-down and didn’t want to interfere
  - found log-normal distributions (long tails) and that lack of information sharing was damaging effectiveness
- Who could we tell anyway?
  - no organised way to report data to banks
- Our NDAs forbade this!
  - take-down companies make money by selling data feeds
- c.f. clinical trials: These trials can and should be stopped prematurely once the results become statistically significant and the divergence in treatment outcome is substantial
- We recommend that researchers avoid direct interference during data collection, but once the conclusions have been drawn, assistance to relevant stakeholders should be encouraged.

Dilemma 2: Should researchers intervene to assist victims?

- We reported on 414 compromised users whose details we recovered from phishing sites (and have found more since)
- We repatriated these to banks where we could (and there is since 2010 a formal scheme for this run by the NCFTA)
- Common issue: Torpig takeover, 180000 infections, 70G data, 1 million Windows passwords, 100000 SMTP logins, 12000 FTP credentials. Disrupted research for 6+ months
- But may be hard to locate victims, BBC’s “Click” changed ‘wallpaper’ on botted machines; apparently without having considered that this is clearly a s1 offence under the Computer Misuse Act 1990.
  - FBI was proactive on Coreflood, but only in the jurisdiction

Rod Rasmussen (Internet Identity)

The normal admin for the machine had been deployed to Iraq as part of his National Guard unit, and his backup was busy and hundreds of miles away that weekend because of his father's funeral. There were plenty of people looking at the machine (as in had their physical eyeballs on it) including the local sheriff, but no one was touching it since it ran the 911 Dispatch system and no one had the knowledge (as in passwords and expertise) to fix it.

We've also had take-downs on machines that were in hospitals, railroad stations, airports, and government facilities. While those could be just public access terminals, there's no way we can tell from the outside if that is the case or they are running life-saving equipment, switching operations, air-traffic control systems, or have sensitive data on them respectively. That's why we have a very bright line barring any sort of "write access", resetting or otherwise monkeying with content on compromised servers. Not only is it usually illegal in the US, someone's life can literally be on the line!

Dilemma 3: Should researchers fabricate plausible content to conduct 'pure' experiments of take-down?

- Most empirical research in computer security is 'observational'
- Some attempts at experiments, eg copyright issues, but considerable flaws since didn't investigate whole process (especially US DMCA 'put-back')
- Risk of wasting the time and energy of frontline responders on fabricated requests suggests real harm is caused by the experiments. In particular, the responders typically have substantial resource constraints and already find it difficult to keep up with the number of legitimate take-down requests
- We believe the fabrication of reports to study take-down is usually unethical

Dilemma 4: Should researchers collect world-readable data from ‘private’ locations?

- We used “Webalizer” data from compromised websites to study the number of victims and how sites were located by criminals
- Owners of these sites may not have intended to publish their visitor data – but they did; so there’s an ethics question
- Our view was that the data enabled us to answer questions that we could not have done otherwise; and it was not personally identifiable data, just summary data for website visits
- On balance, we feel the opportunities for scientific advancement outweigh the risks to an individual website operator in collecting the data. However, it is a judgment call, and one that should be weighed on a case-by-case basis.

Dilemma 5: What if our analysis will assist criminals?

- This is always an issue; the issue is sometimes described as the suitability of ‘full disclosure’
- We observed that fast-flux systems used multiple servers but this is unnecessarily cautious.
- We also made some observations about DNS time-to-live values
- We tried not to emphasis these issues and the criminals have not changed how they operate
- Long history of ethical analysis of this issue, for example Hobbs (1853) re locksmiths, and Wilkins (1641) re crypto
- General view is that the benefits of explaining how the criminals operate benefits the good guys far more than the bad guys, who already know how to do their crimes and already understand what works and what doesn’t

Dilemma 6: Should investigatory techniques be revealed?

- It’s not science if an academic paper does not explain the methodology of an investigation
- The main effect of this has been for us to suppress more minor bits of research
- Others take a different line, such as Netcraft explaining about phishing kit ‘back doors’ and Billy Rios discussing a file injection vulnerability in Zeus
- We don’t agree that ‘full disclosure’ is always the overriding principle and so we choose not to publish when the details of our paper would disrupt investigations by the authorities

Dilemma 7: When should datasets be made public or kept secret?

- Phishtank lists phishing websites, but others (Google, Microsoft, Netcraft and the take-down companies) keep lists private
- We found that sites on Phishtank were less likely to be recompromised & we found that sharing data between take-down companies would reduce phishing website lifetimes
- However, there are downsides to public sharing, it shows what the defenders know, it ‘names & shames’ – and criminals can steal caches of credentials (as we did)
- Many systems hash the URLs so they can be compared, but the actual values are not revealed – this prevents research and prevents defenders being proactive.
- Decisions on publishing are ethical decisions

Dilemma 8: Is the fix realistic, and does it consider the incentives of all the participants?

- We believe in security economics, and so one of the ways we look to solve security problems is to align incentives
- However, when we proposed list sharing we did not take full account of the incentives of the take-down companies; though we fixed this after they complained of our naivety
- We have proposed improvements in the take-down of child sexual abuse, unfortunately they are not realistic whilst INHOPE prevents cross-border notifications. We still think this would be the right thing to do.
- It is unethical to propose fixes to security problems that cannot be made to work in the real world

Dilemma 9: What if the fix is worse than the problem?

- Restrictions on registering domain names would help the problem of misleading names but would not be proportionate, which is why we never suggest that type of solution
- Paul Vixie promotes an efficient way (RPZ) of publishing lists of domains that are to be suppressed by Domain Name Servers (similar to the way that spam senders are currently handled). However, it seems unwise to institutionalise this type of suppression in a world where many groups see removal of domain names as a way to impose their world view
- We think that it is unethical to propose fixes without considering their impact, and seeking to minimise the side-effects

Some more “war stories”

- Publishing the website that was multiply compromised
- Being the subject of a RIPA s22 notice
- Being DDoSsed by the storm botnet
- Being DDoSsed by the owner of a stressor

Ethics paper conclusion

http://www.cl.cam.ac.uk/~rnc1/ntdethics.pdf

- We’re not proposing ethical principles but telling ‘war stories’
- But philosophers might usefully take our stories into account
- Also, if you’re going to work in this area it would be worthwhile learning from our mistakes!

Final section of this talk is about cybercrime data

Big !

Biased ☹

and Beyond Review ?

Real datasets are big(gish)!

- Phishing URLs feeds run at perhaps 750K+ a year
  - if you don't dedupe you will be seriously struggle
    - they are inflated by passive DNS (many names for same host)
    - they are inflated by unique URLs per victim
    - they are inflated by unique URLs per lure (esp Facebook phish)
- I track DDoS attacks ... up to 75m+ events a month (events up to 100K packets) : 500m events for about 10K victims/day
  - you haven't read the paper because I haven't done the analysis yet!
- I tracked a "worm" that spread by Instant Message
  - 55m downloads, certainly >3 m victims (eCrime 2015 paper)
  - note that the dataset for Conficker now exceeds 40TB
- If you don't understand what 'scale' means then
  - you'll get swamped and end up data processing rather than thinking
  - you'll only process a subset the data and hope it is representative

Datasets are biased

- I have co-written papers on High Yield Investment Programs
  - these are Ponzi schemes paying %ages per day
  - BUT all the sites are in English ! Are there sites in Russian ?
- Paper ‘to appear’ (RSN) about ‘booters’ (DDoS for hire)
  - sites are in English & French. Are any Korean ? Japanese ??
- PhishTank holds circa 45% of phish URLs (but all eBay/PayPal)
  - cf IC3 and auction fraud
  - surprising number of papers just use PhishTank data
  - what happens if you try to get “phresh phish” for a toolbar study ?
- Bias is acceptable if you can model it, but can you do that ?
  - Moore & Vasek are studying compromised WordPress systems
    - is this <n> attackers ? (some evidence it will be one search engine!)
  - I found one affiliate spammer who used 5K websites / day
    - but did they buy them from <n> people ?

Datasets are proprietary

- For much of the basic data I work with, I have spent years building the relationships and trust needed to obtain the data
    - that makes it hard for others to work in this space
    - & it’s hard for me to work on things where I don’t have contacts
- Much of the basic data I work with comes to me under NDA
    - sometimes I can’t even say who gave it to me
    - specialist companies want it kept away from competitors
        - and especially not given to customers for free!
    - large companies have obligations under privacy policies and may be concerned about the PR aspects of being linked to eCrime
    - some of the data is personal data (Americans would say PII)
    - everyone is concerned not to tell the criminals what we know
    - everyone is concerned not to tell criminals about insecure sites
- So “open data” isn’t going to be happening in this space

Is this science ?

- My research isn't unique in that almost no work on cybercrime can be reproduced – so can we really call this 'science' ?
  - colleagues who work on operating systems and computer architecture say that they have similar experiences
- Now of course you don't get papers published at prestigious conferences by reproducing results. But perhaps you do if show a better or unflawed approach ? But if you're not working on the same data is it correct to make the comparison ?
  - that's not how the physics or chemistry works as a science
    - fortunately, or otherwise, we only have one universe to measure !
- In practice this all means that there's very few undergrad projects or MSc theses that tackle cybercrime. People don't have the data or know it will take years to collect it and so working in this space looks too much like long-term 'research'

The hard question...

If you think my cybercrime work is flawed then how can you hope to improve on it ?

- I am not allowed to give you my data
- It may take you years to get your own data
- The data may be at such a scale as to swamp you
- The data may have biases that you (and I) fail to understand
- The raw data may be only the beginning of your problems because you may have to build a web scrapers, learn how to fetch whois information (running into access limits), avoid refetching the same thing under another name, deal with retaliation from the criminals etc etc

It's easier to answer "why do so few of us work on cybercrime?"

Perhaps I have an answer...

- I have 5 years funding from the EPSRC to create the
  Cambridge Cloud Cybercrime Centre
- Our approach will be data driven. We aim to leverage our neutral academic status to obtain data and build one of the largest and most diverse datasets that any organisation holds
- We will mine and correlate this data to extract information about criminal activity. We will learn more about crime 'in the cloud', detect it better & faster and determine what forenics looks like in this space (and where appropriate work with LEAs)
- We aim to create a sustainable and internationally competitive centre for academic research into cybercrime....
  - BUT you can play too!

Working with C.C.C.C. datasets

- We want to convert our NDAs into ‘DAs’
- We aim to collect data, add value to it and then make it available to others under one (we hope) simple NDA
- We can’t make the data entirely public (or open) – but we will be making it available to legitimate academics
- We aim to have a ‘catalogue’ of data that you can use in your specialist research without you having to learn all about the web scraping, the whois limits, the duplicated data and so on
- We aim to make it easy to set MSc work in this area knowing that it won’t take two years to get the data together
- We aim to see more **science** by letting people run different techniques on the same data and compare results

Fighting crime isn’t inherently competitive

- We’re going to have a LOT of data
- We aim to work on this at Cambridge – we’ll have world class researchers doing world class work
- BUT we’re also going to make that data available to others
- At the end of the first five years I want to be judged not on how many papers we wrote in Cambridge but how many papers you all wrote because we helped to make that possible
- I also want to see MSc and undergraduate students tackling cybercrime projects and confirming or refuting classic results.
  - besides improving the science – this is a fascinating area and I want others to experience that too and come and join us
- We (and you) will find new ways to prevent crime, to detect and deter criminals – and that’s why society funds our work

Cambridge Cloud Cybercrime Centre

https://cambridgecybercrime.uk

security group blog:

https://www.lightbluetouchpaper.org

my publications:

https://www.cl.cam.ac.uk/~rnc1/publications.html

UNIVERSITY OF
CAMBRIDGE
Computer Laboratory