Cambridge Cybercrime Centre

Richard Clayton
Director

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

San Diego
27th February 2020

My background

- I’ve been looking at online abuse (spam, phishing, malware, DDoS etc) for two decades
- My general approach is data driven (I count things)
- I have obtained many datasets from industry under NDAs and that has underpinned the work I have done (in collaboration with some very smart people)
- BUT this is a long and tedious process, and we’re beginning to realise that no papers in this field can be reproduced (data cannot be shared, results cannot be compared, conclusions cannot be validated)
- This does not really look like science…

Cambridge Cybercrime Centre

- I have 5 years funding from EPSRC (+ some other money)
- Currently 6 of us
  - plus PIs, PhD students, MSc students &c
- We are interdisciplinary

  Computer Science & Criminology & Psychology

  and previously Law
- Our approach is 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 forensics looks like in this space (and where appropriate work with LEAs)

Datasets

- Underground Forums (>> 70m posts)
- Discord & Telegram chats (just getting going, 100's of channels)
- Blog spam (>400K posts)
- Reflected DDoS victims (5+ years data)
- Mirai scanning data (of Cambridge and elsewhere)
- Mirai (etc) malware (since Dec 2016, 175K samples!)
- Email spam (back to 2004, and some from the 1990s!)
- 419 scam emails (> 60K, dating back to 2006)
- Phishing emails (50K plus, over 10 years)
- *NEW* email spam from Abusix (c 3M messages a day)

… plus many datasets from our old papers

Our data is being used…

- 33 signed up research groups (~100 researchers)
  - 12 UK, 5 US (4 continents, more in pipeline)
- Most popular dataset is CrimeBB
  - Mirai / DDoS data also becoming more popular
- We're looking hard at how people use our data, how we can make it easier for "ologies" and non-tech people
  - CrimeBB being used by criminologists, sociologists etc. and they can't necessarily cope with SQL databases
  - also, we want to help people learn if we have relevant data for their research projects before they sign the paperwork
  - we want to do more "AI" to label data (and help others do their own labelling and share that) – comparing labelling important in it's own right but also assists in research by identifying active participants

https://www.cambridgecybercrime.uk/process.html

UNIVERSITY OF CAMBRIDGE

Search

Contact us | A–Z | Advanced search

Computer Laboratory

Cambridge Cybercrime Centre: Process for working with our data

This page sets out the steps in the process for obtaining data from the Cybercrime Centre.

Assess whether you will be allowed to use our data

Our datasets are intended for research and analysis into methods to find, understand, investigate and counter cybercrime so your project must clearly fall into this space. Although we do not require reseachers to be academics, there are significant restrictions on using our data for commercial purposes.

Although some of our data was generated internally and so we can make it available for other types of project and for commercial purposes, much of our data has come from third parties and they have only provided us with the data because of the framework under which it will be shared.

Identify the data you wish to use

We describe our various datasets on this page [LINK]. The descriptions are public and necessarily fairly high level. We do however try to indicate the size of the datasets, the period over which they was collected, along with any known biases.

We strongly enourage the use of prepacked datasets rather than "live feeds". Although a live feed may be superficially attractive it makes it harder to arrange that other researchers can receive the same data that you did -- a key aim of the Cybercrime Centre is to enable reproducible research. If the issue is that you need to collect a further "field" over and above what we supply then talk with us and we may well be able to do this for you.

Read about our legal framework

It is important that you understand the basis on which we share data and the paperwork that will need to be signed.

There's several pages of explanations and FAQs about our agreements, starting here at https://www.cambridgecybercrime.uk/data.html, which you should read before contacting us.

Make an application

You will need to make a formal application to use our data. In the first instance you should send an email to the Director of the Cybercrime Centre,

Where does our data come from ?

- Original idea was to use my connections with industry
  - but this has proved difficult (and no-one interested in phishing)
  - AbuseIX is a key exception … and we will see how that goes
- Most data we collect ourselves
  - post-docs are expected to spend a lot of time collecting new data
  - the secret sauce is to implement "production" systems to collect the data rather than ad hoc collection for a particular purpose
- I share data NOW (real time if necessary)
  - if we haven't looked at the datasets yet, then more fool us
- Data is essentially all public which simplifies the legalities
  - however, I sniff JANET traffic which is lawful because it keeps Cambridge safe BUT I cannot legally share raw traffic
    - note that our ethics case only permits examination of incoming conversations (and never email)

Legal framework

- We share data under an NDA (technically it's a license)
  - executed between Cambridge and your institution
  - "incoming" licenses are with Cambridge and allow us to share data under the standard "outgoing" framework
- Purpose MUST be to tackle cybercrime
  - "incoming" and "outgoing" have to match!
  - BUT where data entirely ours we might share anyway
- I am not very flexible about outgoing terms but very flexible about incoming data
  - in particular about publicity (or otherwise) for you
- I want to be a one-stop shop for sharing with academics
  - viz: I'll handle your data for you
- GDPR is not an insurmountable barrier!

Outcomes

- People are doing research with our data
  - yay!
- People I've never heard of are doing research with our data
  - this is the most cheering aspect of what we are doing
- People are writing papers using our data
  - 17 papers in our list and more in prospect (it takes a long time for papers to appear!)

Funding and the future

- Initial 5 year EPSRC grant ends in September
- We expect to press on at a reduced level using ad hoc grants (and donated effort) to keep systems running
  - existing emphasis on "production" systems means ongoing effort is not outrageously high – impact will be mainly on identifying new types of data to collect and building new collection systems
  - we have a fair number of servers and spinning disks, so capital costs are low in the short to medium term
- UK Research Councils not oriented towards funding "infrastructure" (if it is not a space telescope or similar)
  - we think that "infrastructure" is the right analogy, but funding this on either a national or international basis is an unusual ask at the present time
- There's also challenges in funding interdisciplinary teams

Other sharing regimes

- IMPACT
  - more like eBay (CCC more like Amazon !)
  - because only the vetted can browse can describe data better
  - once you have found a seller then negotiate terms
- APWG
  - has shared phishing URLs for 16 years
  - has branched out into other threat indicators (bitcoin wallets, malware, VPN connections &c)
  - easy for academics to get access (once you know about it)
  - data not especially “clean” (so you need to remove rubbish)
  - APWG pioneering a GDPR Section 40 (“Codes of Conduct”) approach to sharing data (and in particular IP addresses)
    - very formal; involves mandatory monitoring / auditing; BUT should be very useful once in place (it’s a cutting edge WIP at present!)

https://cambridgecybercrime.uk

our blog:

https://www.lightbluetouchpaper.org

UNIVERSITY OF CAMBRIDGE
Computer Laboratory