Cambridge Cybercrime Centre

Richard Clayton
Director

Cambridge
13th July 2017

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 6.5 of us
- We are interdisciplinary

  Computer Science, Criminology & Psychology
- 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)

Others can play too

- We have started the process of renegotiating existing NDAs
- We will collate the data, add value to it, put it into collections; and then make it available to others under one simple NDA agreement which is between the researcher and us
- We cannot make the data entirely public (or open) but we *will* be making it available to legitimate academics
- We will have a 'catalogue' of data that can be used in specialist research without the need to learn all about the web scraping, whois limits, duplicated data and all the other complexity
  - it will be easy to set MSc work in this area since it will not take 2 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

This is not a competition

- We will use this data in Cambridge – we will have world class researchers doing world class work
- BUT 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 were written across the whole of academia because we helped to make it possible
- We (and the people using our datasets) will find new ways to prevent crime, to detect and deter criminals – and in the end, that's why society funds our work

Datasets

- Phishing emails (60K plus, over 10 years)
- Phishing URLs and pages
- Underground Forums

HackForums	(19m posts, 2.6m threads)
Offensive Community	(57K posts, 4.5K threads)
Kernelmode	(26K posts, 2.8K threads)

- Blog spam (>150K posts)
- Reflected DDoS victims (3+ years data)
- Mirai scanning & malware (since Dec 2016, 6200 samples!)
- 419 scam emails (> 75K, dating way back)
- Email spam (back to 2004, and some from the 1990s!)
- SSH honeypot datasets (> 1 year)

… plus many datasets from our old papers

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,

Join in!

https://cambridgecybercrime.uk

our blog:

https://www.lightbluetouchpaper.org

UNIVERSITY OF CAMBRIDGE
Computer Laboratory