Cambridge Cybercrime Centre

Richard Clayton
Director

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

Cambridge
11th July 2019

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 (and I'm currently recruiting 1 more)
  - 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 (>> 60m posts)
- Discord & Telegram chats (just getting going, 100s channels)
- Blog spam (>400K posts)
- Reflected DDoS victims (5+ years data)
- Mirai scanning data (of Cambridge and elsewhere)
- Mirai (etc) malware (since Dec 2016, 120K samples!)
- SSH honeypot datasets (> 3 years)
- 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)

… plus many datasets from our old papers

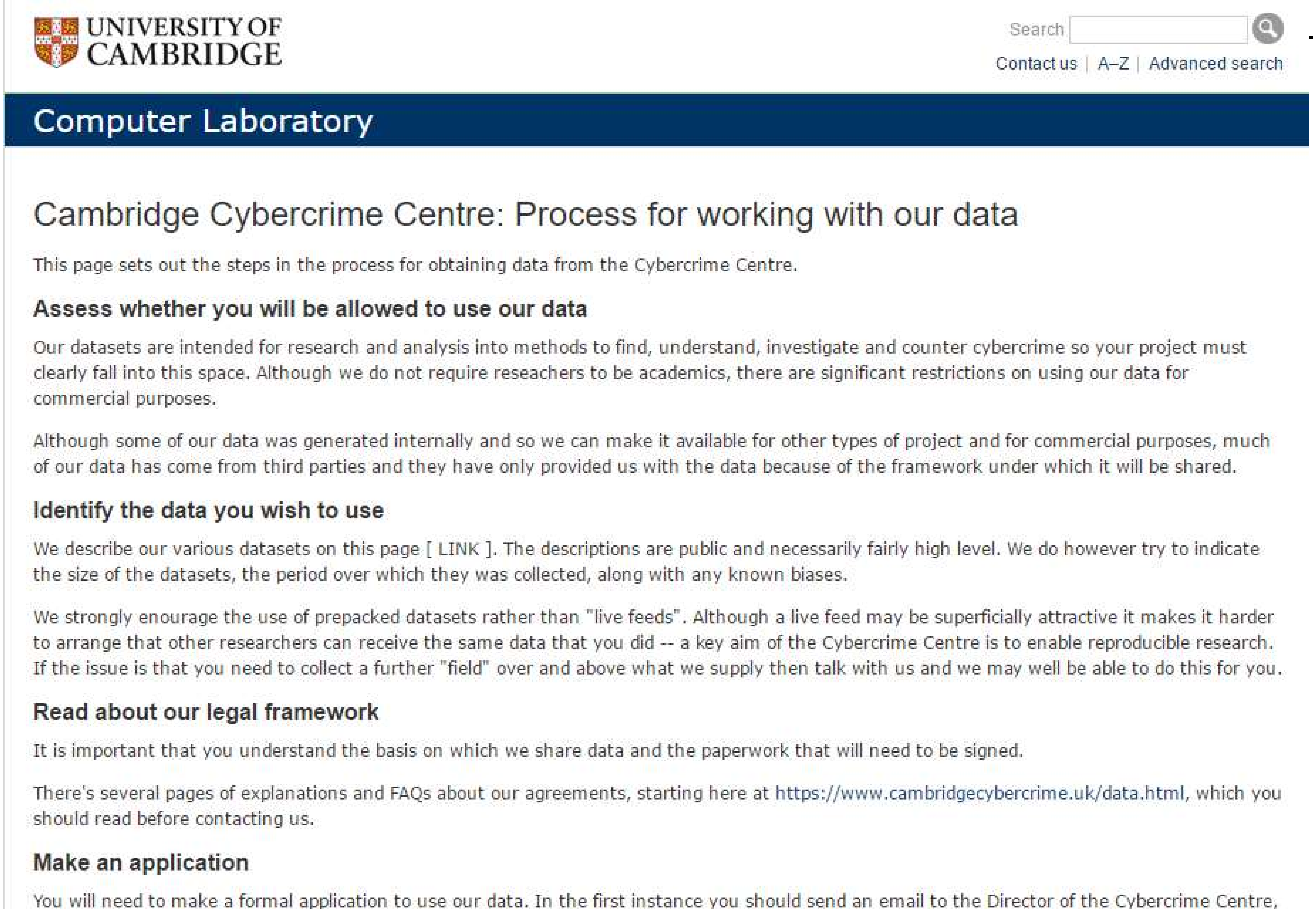
Our data is being used…

- 26 signed up research groups (> 50 researchers)
- 5 continents!
  - 10 UK, 4 US

- Most popular dataset is CrimeBB

- We’re looking hard at how people use our data, how we can make it easier for “ologies” and non-tech people
  - also, we want to help people learn if we have relevant data for their research projects
  - we want to do more “AI” to label data (and help others do their own labelling and share that)

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



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`

come back next year: July 2020

UNIVERSITY OF
CAMBRIDGE
Computer Laboratory