Measuring & Studying Cybercrime

Richard Clayton
Cambridge Cloud Cybercrime Centre

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

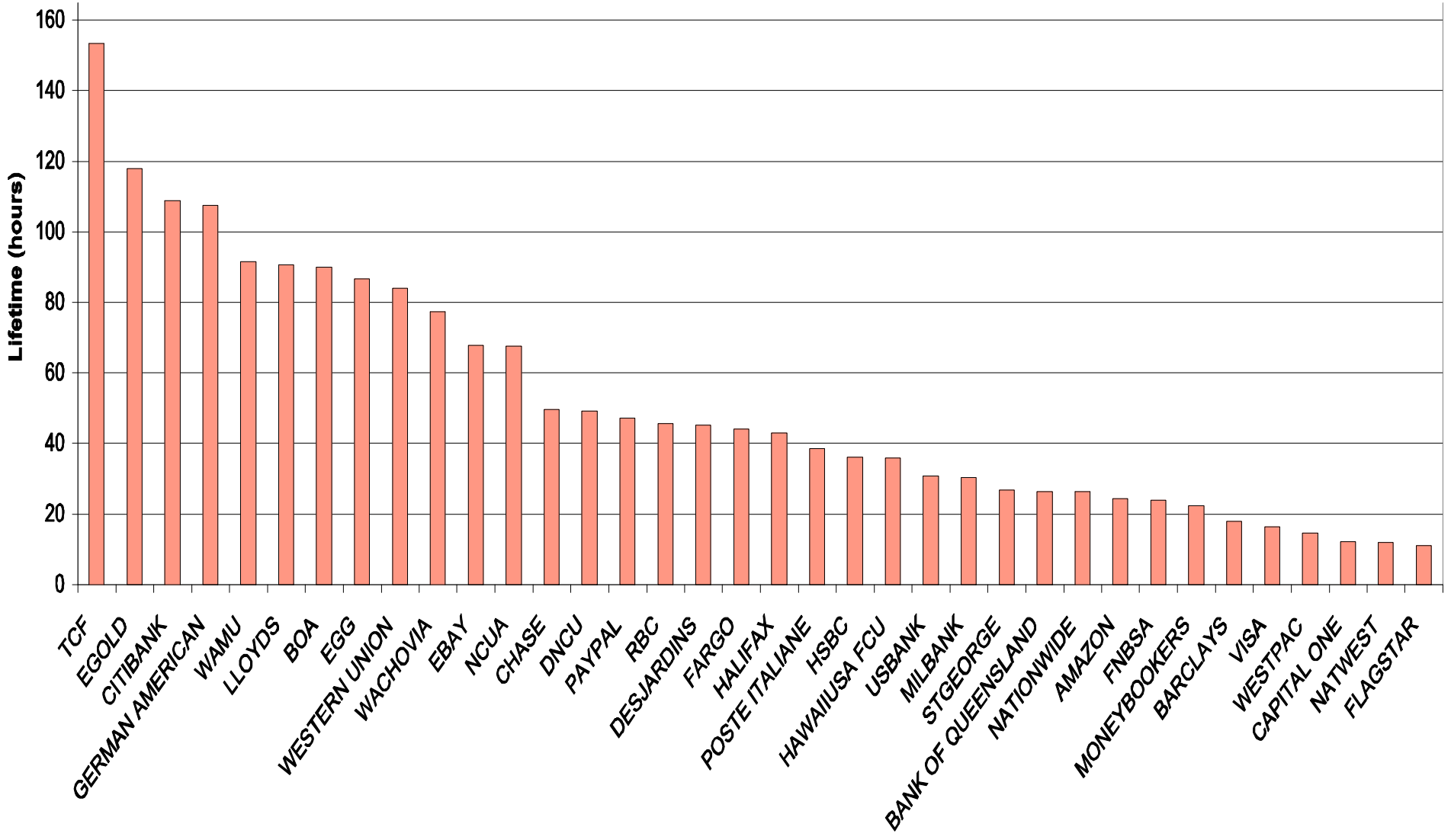
Cambridge
14th July 2016

Phishing research (with Tyler Moore)

- In 2007 Tyler and I looked at phishing (fake bank logins)
- Our main innovation was to measure website lifetime so we built an infrastructure to visit dubious URLs every 30 mins…
- We were lucky in that just as we started some of the criminals [the "rock-phish gang"] started using "fast flux" – their domains resolved to HTTP relays (a different set of N every few minutes) and the relays then connected to a hidden "mothership"
- The only viable defence was to "take-down" the domain name rather than the individual relays
- So we were able to contrast take-down times to see if the criminals actually had an edge…

Phishing website lifetimes (hours)	# sites (8 weeks)	Mean lifetime	Median lifetime
Non-rock	1707	58.4	20
Rock-phish domains	419	94.3	55
Rock-phish IP addresses	122	124.9	25
Fast-flux rock-phish domains	67	454.4	202
Fast-flux rock-phish IP addresses	2995	124.6	20

We looked at take-down time per brand



Reaction to our research

- The take-down industry was very interested
  - BUT they said our average measurements were too high
  - take-down was hours not days
- We explained about "long tails" etc.
  - and how careful we'd been and how clever!
- They said we should try again with more data
  - and they gave us more 'feeds' of URLs
- AND then one day Tyler came to see me and said he understood what was going on…
  - the companies did not share data with each other, but they all shared with us
  - we had learnt about Bank X sites, but Bank X's provider did not know about them yet SO they did not do start any take downs …

Our January 2008 data

	Total	Mean (hours)	Median (hours)
Free webhosting	395	48	0
when brand owner aware	240	4.3	0
when brand owner unaware	155	115	29
Compromised machines	193	49	0
when brand owner aware	105	3.5	0
when brand owner unaware	88	104	10
Rock-phish domains	821	70	33
Fast-flux domains	314	96	25

… and the key lessons are

- Datasets are bigger than you think
  - a few thousand sites was a struggle
- Datasets contain lots of errors
  - many “phish” were something else
- Datasets are biased
  - but this can lead to more understanding
- Datasets are proprietary
  - sorry, I cannot share this data with you!

Datasets are big!

- Phishing URLs feeds run at perhaps 750K+ a year
  - must dedupe irrelevant parts of URL
  - must remove effect of passive DNS inflation
  - must deal with URLs being unique per victim
  - must deal with URLs being unique per lure
- DDoS attacks … up to 75m+ events a month
  - in two years we've analysed 1.25 trillion packets
  - around 5000 victims per day
- If you don't understand "scale" then:
  - you'll get swamped and end up just doing data processing rather than thinking
  - OR you will only process a subset the data and hope that it is representative

Datasets are biased

- PhishTank holds c 45% of phishing URLs
  - but it has c 100% of all eBay/PayPal phish
  - if you don't know that, you will be misled
  - there's a similar issue with IC3 and auction fraud

- Bias can work for you
  - BUT only if you can model it; but can you always do that ?
  - e.g. an affiliate spammer is using 5K websites/day
    - but did they buy them from <n> people ?
    - are you studying an ecosystem or just one person ?

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
  - it is hard to begin working in this space
  - AND it is hard for me to work on topics where I do not have good contacts

- I receive data under Non-Disclosure Agreements (NDAs)
  - and sometimes I cannot even say who the NDA is with

“Open Data” is not possible

- Companies will give me (and other academics) data:

BUT they may not want publicity about ecrime

AND I must not give data to competitors

AND it must not be given to paying customers

AND privacy policies may impose restrictions

AND there may be personal data (or PII)

AND we must not tell criminals what we know

Is this “science” ?

- My research is not unique, in that almost no work on cybercrime can be reproduced – so can we really call this “science” ?
  - now of course you will not get papers published at prestigious conferences by reproducing results. But perhaps you can if you show a better approach or find flaws in the original paper? But if you are not working on the same data is it correct to make the comparison?
- Very few undergrad projects or MSc theses tackle cybercrime, but in many research fields many MSc projects (attempt to) reproduce earlier work
  - people don’t have the data, or know it will take years to collect it, and so it looks too much like long-term “research”
  - SO we don’t check that classic results are correct
  - AND we do not get young researchers interested in cybercrime

Let’s have more research !

- We’d like to see more cybercrime research, BUT:

  I am not allowed to give you my data

  It may take you years to get your own data

  The scale of the data may swamp you

  We may all fail to understand bias in the data

- And data is just a start: web scraping, whois, etc.
- The easiest question to answer is:
-

  “why do so few of us work on cybercrime”

I have an answer

- I have 5 years funding from the EPSRC for the

Cambridge Cloud Cybercrime Centre

- We aim to create a sustainable and internationally competitive centre for academic research into cybercrime….

Our approach

- 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 forensics looks like in this space (and where appropriate work with LEAs)

You can play too

- We have started the process of renegotiating our existing NDAs
- We will collate the data, add value to it, put it into collections; and then make it available to others under one (we hope) simple NDA which is between the researcher and us
- We cannot (see earlier) 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 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
  - 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 you all wrote because we helped to make it possible
- We (and you) will find new ways to prevent crime, to detect and deter criminals – and in the end, that's why society funds our work

Join in!

https://cambridgecybercrime.uk

our blog:

https://www.lightbluetouchpaper.org

my publications:

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

UNIVERSITY OF CAMBRIDGE
Computer Laboratory