Cybercrime data: Big, Biased and Beyond Review?

Richard Clayton

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

ESORICS
23rd September 2015

Outline

- The cost of cybercrime
- A story about phishing
- What do I mean by big ?
- What do I mean by biased ?
- What do I mean by beyond review ?
- Why is it hard to research cybercrime ?
- Why I think our new initiative will help!

Measuring cybercrime

- 2009 McAfee: cybercrime costs $1000bn ($1 trillion) worldwide
- 2011 Detica (part of BAE plc): estimated cost of cybercrime to the UK economy was $43 billion / annum (~ 1.8% of GDP)
  - a main contribution was "industrial espionage"
- Florencio and Herley "Sex, Lies and Cybercrime Surveys"
  - this WEIS 2011 paper points out how outliers affect results (single loss of $50K in a 1000 person survey becomes $10bn scaled up)
- Inga Beale (CEO Lloyds of London) Jan 2015 said cost of cybercrime to companies was $400 billion a year… or did she?

Fortune: "*Beale said that Lloyd's estimates that cyber attacks cost businesses as much as $400 billion a year, including the damage itself and subsequent disruption to the normal course of business*"

The cost of cybercrime

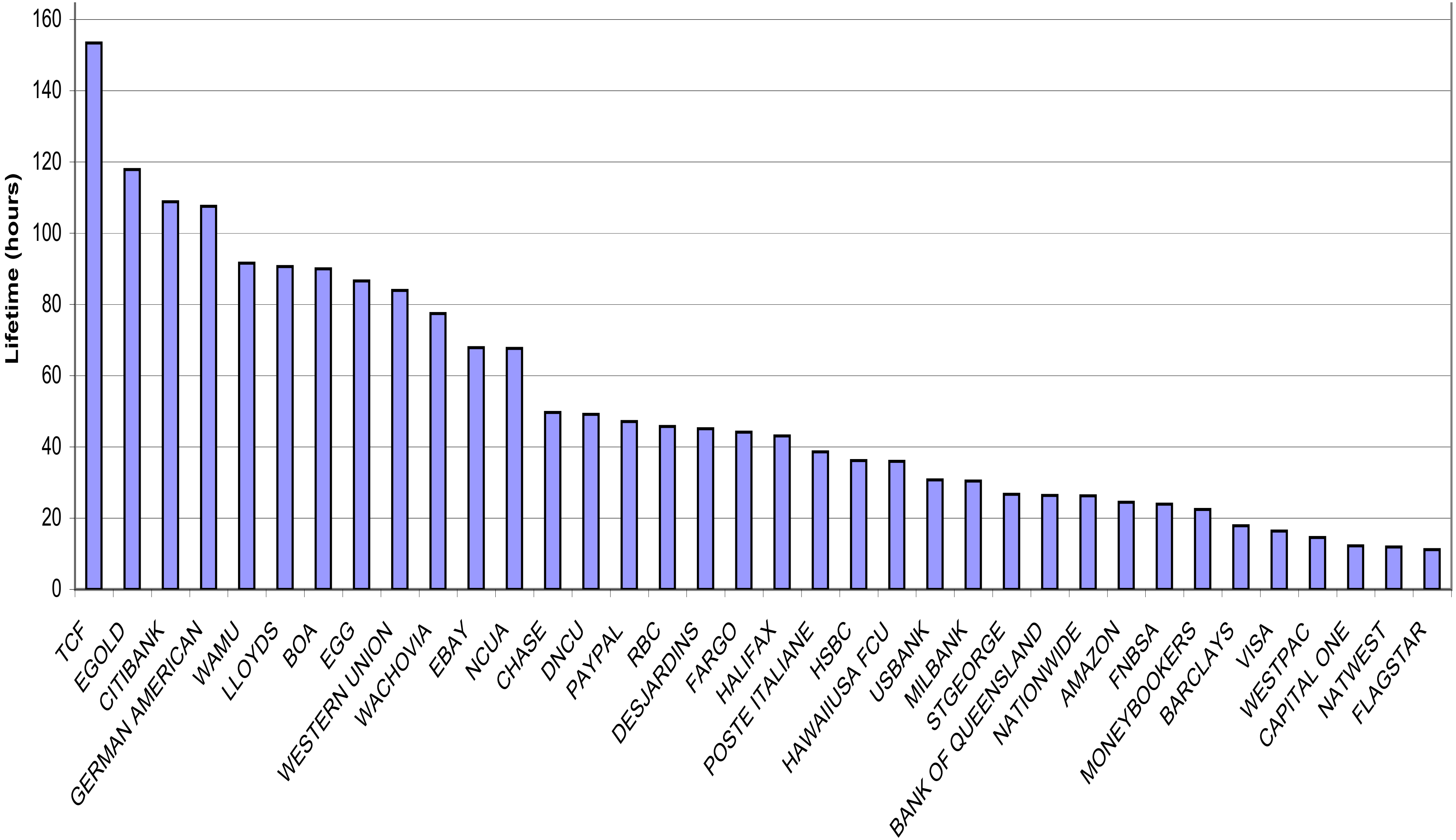
- There's a number of academics writing papers on different types of cybercrime, phishing, unlicensed pharmacies, extortion etc.
- Police publish data about some of the crimes they deal with
- We (multiple expert authors) assessed data for WEIS 2012
  - created framework, and gave best estimates for each category
- It didn't add up to as much as you'd think:
  - traditional frauds cost citizens a few hundred dollars per year
  - transitional frauds cost citizens a few tens of dollars per year
  - new cybercrimes net criminals tens of <u>pence</u> per citizen per year
- BUT the indirect costs and defence costs (& especially clean-up costs) for new crimes are more than 10x the criminal revenue and so Beale is on the right track... (but I still think her figure is absurdly high when measuring criminality)

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



We were given more data...

- The take-down industry was very interested in our research
- BUT they said our average measurements were too high
  - take-down was merely hours they said, not days
  - we explained about “long tails” etc
    - and how careful we’d been and how clever!
  - they said we should try and again with a better set of data
- SO we processed feeds of data from even more companies
- 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 knew of Bank X sites that Bank X’s provider didn’t know about
  - so they didn’t do any take down...

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

So what do we learn from this story ?

- Datasets are big
  - the few thousand sites we looked at were a struggle to cope with
- Datasets contain lots of errors
  - many sites that were said to be phish were something else
- Datasets are biased
  - our first dataset was a mixture of public data and data from one company. They used our graph for years to show that the banks they provided take-down services to had lower lifetimes (well duh!)
- Datasets are proprietary
  - I cannot give you data (from any of these studies) for you to check we added it up right, or to allow you to combine it with your data.
    - I signed NDA’s (& cannot even disclose some sources)
  - we did our best by describing our datasets (so a savvy reviewer could have rejected our first paper… but no-one knew any better)

Real datasets are big!

- 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 in submission 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" ?
  - my 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 "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

Join in! (we start on 1st October)

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