What we now know about phishing websites

(summer 2009 version)

Richard Clayton

(joint work with Tyler Moore & Henry Stern)

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

Google
10th August 2009

What is phishing?

- Capture of user credentials through impersonation
  - in 1996 this was pretending to be an AOL sysop
  - since 2003 has been the creation of fake bank websites
- “Bank” is merely generic – attackers impersonate auction sites, payment processors, Adwords, Habbo, IRS etc, etc
  - common theme is that credentials are worth money
- Losses often quoted as over $2 billion/year
  - loss figures are (dubiously) scaled up from phone interviews
  - Some of Gartner’s figures included lottery fraud scams
  - UK banks lost £53 million in 2008 (£20m-30m in previous years)
- Phishing rare in (eg) Germany – attacks are mainly keyloggers
- Some markets use 2-factor (TANs, CAP, SecureID etc)
  - just means that attacks must be done in real-time

Academics & phishing

- Everyone can play! Display instant expertise!!
  - examine psychology, attempt to block spam, detection of websites, browser enhancements, password mangling, reputation systems etc
- Our approach : Security Economics
  - phishing will continue, so we measure impact, assess the effectiveness of countermeasures, aim to work out how to change incentives so that problem tends to fix itself…
- Hard to report on an on-going understanding
  - papers have to be “novel research”, PhDs have to be “a contribution” – so we pick the “low hanging fruit” and move on
- Errors in early papers often go uncorrected
  - “peer review” process needs knowledgeable peers
  - natural tendency not to want to report failures
  - natural tendency not to admit mistakes

Types of phishing website (Jan 2008)

- Misleading domain name (unusual at present)

  http://www.banckname.com/

  http://www.bankname.xtrasecuresite.com/

- Insecure end user or machine (76% of sites)

  http://www.example.com/~user/www.bankname.com/

  http://www.example.com/bankname/login/

- Free web hosting (17% of sites)

  http://www.bank.com.freespacesitename.com/

- Specialist attackers
  - distinctive patterns, often rely on wildcard DNS
  - figures only meaningful after canonicalisation
  - rock-phish 4%, fast-flux 1.4%, “ark” 1.4%

Rock-phish & fast-flux mechanisms!

- Rock-phish (originally used /rock then /r1)
  - compromised machines run a proxy
  - domains do not infringe trademarks
  - name servers usually done in similar style
  - distinctive URL style

    http://session9999.bankname.com.lof80.info/signon/
- “fast-flux” appeared in Feb’07, exclusive since July 08
  - also uses proxy machines that relay “mothership” traffic
  - hostname resolves to 5 (or 10…) IP addresses at once
  - BUT in 20 minutes time, resolves to a different set of machines
  - name server operates in the same way
- Tackling these sites means suspending the domain name, because cannot tackle the proxies fast enough

Take-down time measurements (Jan 2008)

	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	155	104	10
Rock-phish domains	821	70	33
Fast-flux domains	314	96	25

Why are brand owners “unaware”

- Most brand-owners outsource take-down to specialist “brand protection” companies
- These companies compete not only on removal times, but also on how many websites they know of (“the quality of their feed”)
- They get data from “industry” lists (APWG etc) and also from their own spam-traps (old domains, honeypots etc)
- So if Bank X hires company A, but only company B knows about the phishing site then it isn’t removed
- However, as neutral academics we get data from both A and B, we know of the site and measure its (rather slow) removal
- We recommend industry-wide data sharing; the companies buying services from the competition as well!

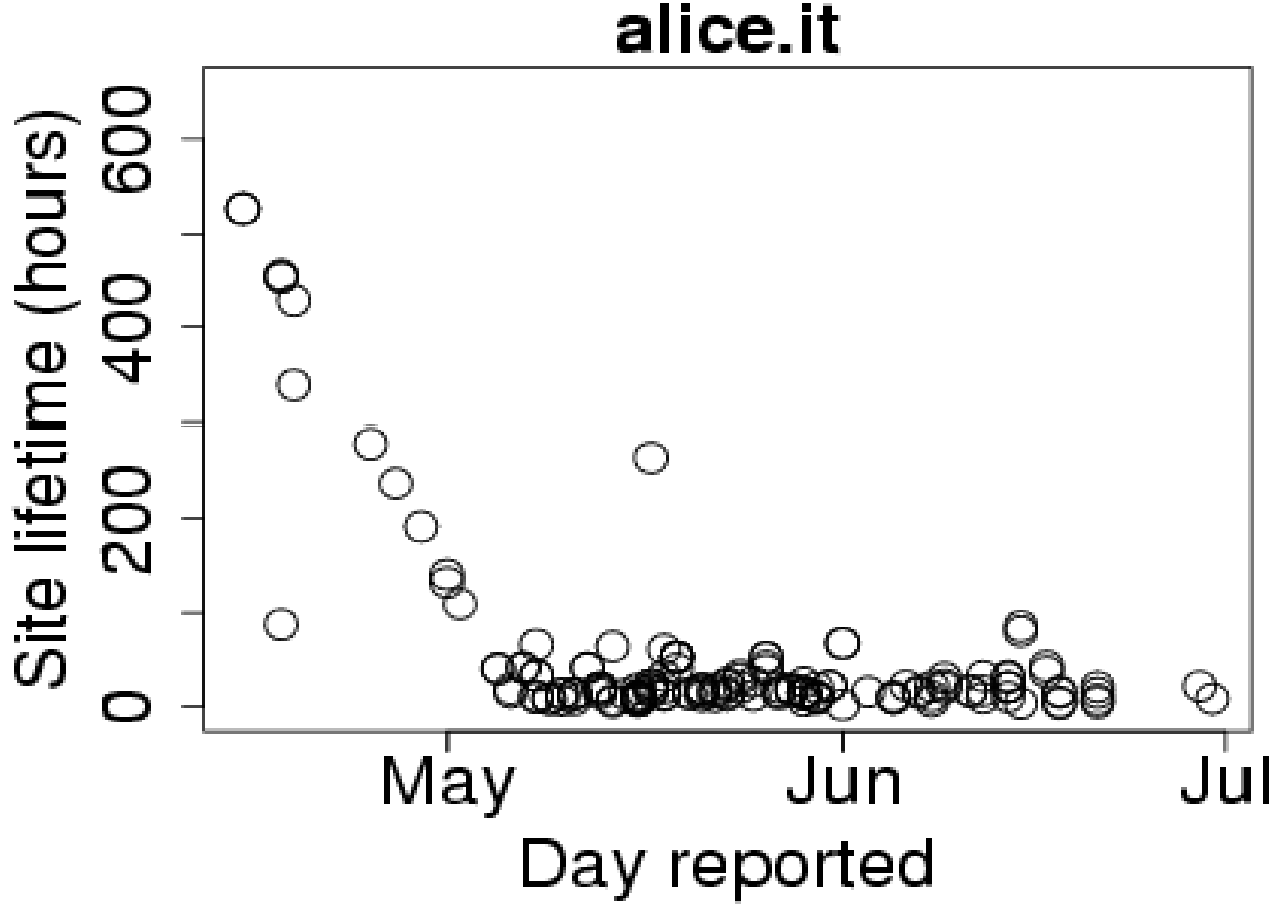
Free web-hosting take-down data (Spring 2007)

Site lifetime (in hours)	# sites	mean	median
yahoo.com	174	23.8	6.9
doramail	155	32.8	18.1
pochta.ru	1253	33.8	16.8
alice.it	159	52.4	18.8
by.ru	254	53.1	38.2

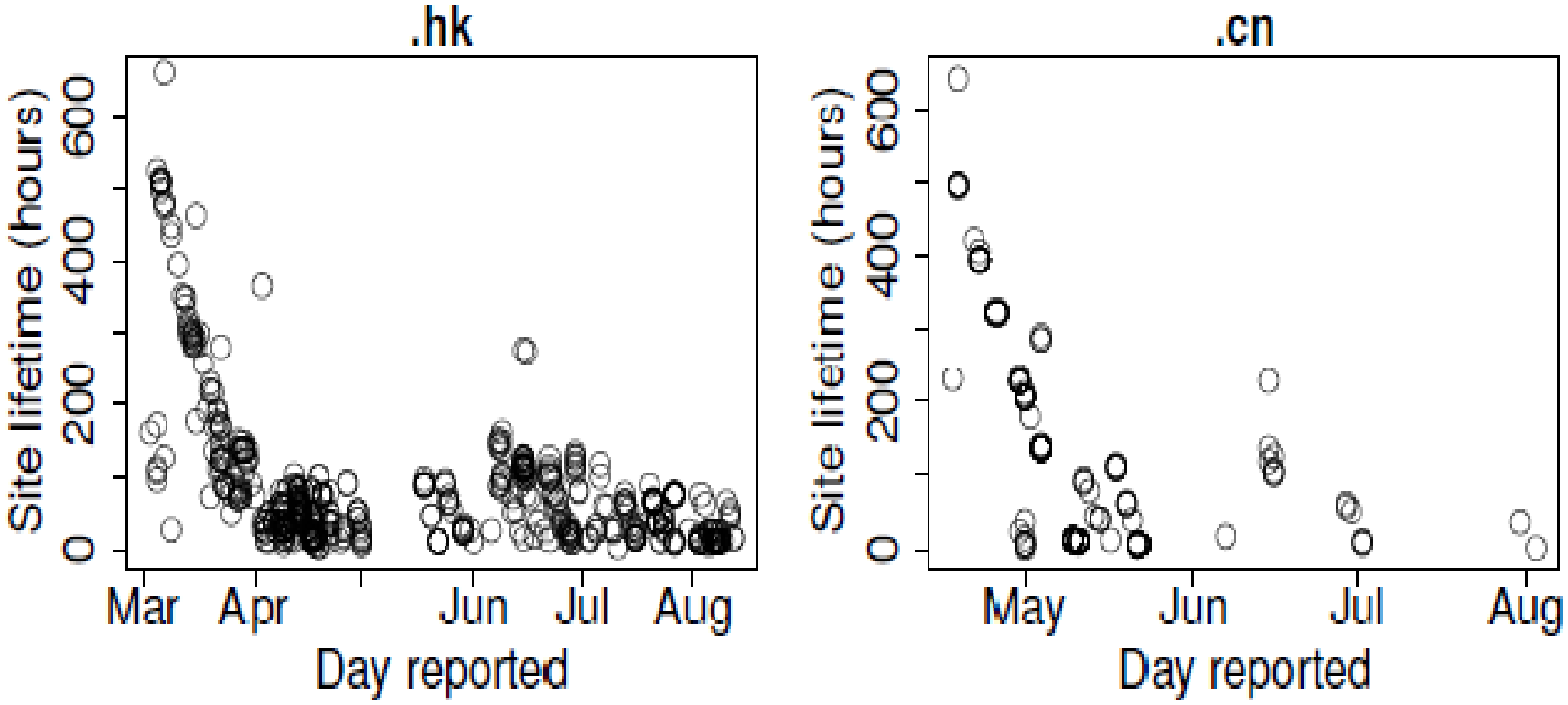
BUT interpret this data carefully: almost all sites (except on Yahoo!) were eBay (65 hour average; this is 1/3 of their total)

ALSO these figures do not account for “awareness”

The gaining of “clue”



Registrars can also have a “clue” issue



How many visitors?

- Some (non rock-phish) sites had world-readable “webalizer” statistics pages which we checked for phishing page visits
  - could determine number who filled in the forms each day
  - 22 on day first reported, 24 next day (then less, but NOT zero)
- Some sites had world readable files of compromised credentials
  - about 50% were “die spammer die” responses
- Hence able to do a sum (Spring 2007 figures)
  - 56 days, 1448 banking websites (exclude eBay)
  - Average lifetime was 57 hours, hence 33 real victims per site
  - Gartner loss estimate of $572/victim (from a $2 billion total)
  - Hence $178 million per year
  - These sites are 1/3 the spam… so $500 million
  - NB: complete hand-waving !!!
  - BUT: is 0.34% of US users; *cf* Florêncio/Herley 0.40%

Non-$hared info al$o repreent ri$k

- Longer lifetimes => more visitors (Webalizer logs)
- Hence we can assess impact of longer lifetimes:

Exposure figures (6 month totals)	A's banks		B's banks	
	K hour	$ million	K hour	$ million
Actual values	1005	$276	78	$32
Expected if sharing	418	$113	61	$28.5
Effect of no sharing	**587**	**$163**	**17**	**$3.5**

- Don't use this table to select a take-down company !
  - A's clients are mainly large banks where lots of phishing sites exist; however, B's clients are smaller and have very few attacks.

How are insecure machines found?

- Traditionally machines found by “scanning” hence interest in Intrusion Detection Systems, “slow scan” software etc etc
- But the Webalizer also parses referrer strings to determine the search terms used to locate the sites...
- Hand categorisation of terms, but most were obvious
  - many searches for MP3s in the logs ! these were ignored
- Types of searches:
  - Vulnerability
    - **`phpizabi v0.848b c1 hfp1`** (CVE-2008-0805)
  - Compromise
    - **`allintitle:welcome paypal`**
  - Shell
    - **`c99shell drwxrwx`**

Webalizer logs (June 07 – March 08)

- 2486 domains with world-readable logs
  - 1320 (53%) had one or more “evil” search terms (they are sometimes called “googledorks”)
- 25 cases where we had sufficient data to prove that searches were linked to the compromise

	Domains	Phrases	Visits
Any evil search	204	456	1207
Vulnerability search	126	206	582
Compromise search	56	99	265
Shell search	47	151	360

Recompromise

- Consider phishing pages on same site more than a week apart (likely a different attacker)
- 9% of all sites recompromised within 4 weeks, rising to 19% within 24 weeks
- For Webalizer sites this is 15% rising to 33%
- If evil search terms present then this becomes 19% rising to 48% (14% to 29% if no terms)
- This doubling is statistically significant!
- The “take-home” from this is:
  - independent attackers are using “search” and finding the same sites
  - websites are being cleaned, but the underlying problem isn’t fixed

Must consider email spam data (Sep 08)

- Email drives visitors to phishing websites
  - assuming equally convincing, this means that losses to customers will correlate closely with spam volumes
- Considered all new sites 24–30 Sep 2008
  - 4084 websites (compromised & free hosting), 120 fast-flux domains
- Matched (generic) URL to an email dataset from IronPort
- Limited spam coverage (surprisingly!?!)
  - 430 sites (11%), 103 fast-flux domains (86%)
- “Fast flux” websites had spam campaigns that corresponded very closely in length to website lifetime
- Other websites sometimes advertised weeks beforehand, and for several days after removal (!)

What's doing the most harm?

- Rapid removal of website will of course mitigate impact
- Total lifetimes split 2:1 between "other" and "fast flux"
- Total spam split 1:2 between "other" and "fast flux"
- Number of websites (and amount of spam) affects public perceptions, possibly eroding trust in eCommerce generally
- But sheer volume of spam may better correlate to total losses

	Websites		Lifetime (hrs)		Spam volume
	Total	%	Total	%	
Ordinary	4084	97%	20603	68%	32%
Fast-flux	120	3%	9674	32%	68%

Comparing take-down times

- Defamation – believed to be quick (days)
- Copyright violation – also prompt(ish)
  - experimentally “days”
  - albeit with prompting, suggesting perseverance matters
- Fake escrow agents
  - average 9 days, median 1 day
  - note that AA419 aware of around 25% of sites
- Mule recruitment sites (Sydney Car Center etc)
  - average 13 days, median 8 days
  - doesn’t attack any particular bank, so they ignore the issue
  - slower than escrow sites (vigilantes more motivated ?)
- Fake pharmacies
  - no vigilante groups – so lifetime is ~2 months

Child sexual abuse images (CAI)

- Provided with anonymised data by IWF
  - Jan–Dec 2007 there were 2585 different domains
  - ignoring 8 (free-web?) domains with >100 reports
- Computed initial take-down time (ignored recompromise)
  - mean 21 days, median 11 days
- If include sites with no removal at all
  - mean 30 days (and growing), median 12 days
- Fast in UK : IWF checks with police and then contacts the ISP
  - but “not authorised” to act internationally
  - passes data via UK police to foreign forces
  - also pass to another INHOPE member
- Confusion of aims (removal/catch criminals)

At present...

- The phishing site take-down industry is putting significant funds at risk by not co-operating with each other
- The police are chasing the right gang!
- Search engines are widely used to find websites to compromise (and re-compromise)
- Takedown times quite clearly affected by “incentives”
- Slowness of removal of CAI is a scandal
- We still don’t know “how many phishers are there ?”
- We still don’t know “is this their day job ?”
- We still don’t know “what’s the best way to disrupt phishing ?”
- We still need better data to improve our understanding!

What we now (Summer 2009) know about phishing websites

BLOG: http://www.lightbluetouchpaper.org/

http://www.cl.cam.ac.uk/~rnc1/
http://people.seas.harvard.edu/~tmoore/

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

UNIVERSITY OF CAMBRIDGE
Computer Laboratory