How hard can it be to measure phishing?

Tyler Moore[*]
tyler.moore@seas.harvard.edu

Richard Clayton[†]
richard.clayton@cl.cam.ac.uk

Abstract

Measuring cybercrime might be thought to be easy; if only the criminals and victims would cooperate and provide the data. This short paper explains, by reference to 'phishing' criminality, how even when a great deal of data is available it is far from clear quite what should be measured. Examples show that various groups have, to their own advantage, selectively chosen what they will measure and that failures to fully understand the data collection process can lead to significant bias and the drawing of entirely misleading conclusions.

There are two possible outcomes: if the result confirms the hypothesis, then you've made a measurement. If the result is contrary to the hypothesis, then you've made a discovery. Enrico Fermi

1 Introduction

Phishing is the criminal activity of enticing people into visiting websites that impersonate the real thing, to dupe them into revealing passwords and other credentials, which will later be used for fraudulent activities. Although a wide range of companies are attacked, the targets are usually financial institutions; hence for simplicity, we will describe them as 'banks'.

There is, compared to most areas of cybercrime, a wealth of information about phishing. Public websites such as phishtank.com track the URLs of phishing websites, and considerably more data is available privately from the 'take-down companies' hired by the banks to get the fake websites removed. The Anti-Phishing Working Group (APWG) regularly publishes summary statistics (based on an industry-wide 'feed' of suspicious URLs), and the banks in France and the UK publish summary data about how much money is being stolen.

However, this data is partial – not only in the sense that it is incomplete, but also in that there is considerable self-interest in the choice of measurements. In Section 2 we explain the reasons behind some of these choices, and explain how it is inherently necessary to decide on the policy reason for measurement before deciding what to actually measure. This is not the only problem with measurement, and in Section 3 we explain how a failure to understand data sources led to inaccurate conclusions in a peer-reviewed (and indeed prize-winning) academic paper.

[*] Center for Research on Computation and Society (CRCS), Harvard University, Cambridge, MA, USA.
[†] Computer Laboratory, University of Cambridge, Cambridge CB3 0FD, UK.

February 22, 2010

2 What's a reasonable measurement of phishing activity?

Back in 2003 when phishing attacks on banks first started, it was easy to measure what was going on. The criminal set up a website with a misleading domain name (or broke into an existing website); obtained a list of open mail relays, and sent out an enticing email (in broken English); and waited for the misled to divulge their passwords. The number of 'attacks' could be equally well measured by counting the URLs, or by counting email Subject lines.

Fast forward through seven years of evolutionary pressure and phishing – by the most organised gangs – is now very different. Each email (now with impeccable grammar) is seldom the same as the next, as the criminals seek to evade spam filters. The mail is sent from a constantly changing set of botnet machines so that blocklists are ineffective. Every URL is unique, because spam filters check URLs specially. The phishing websites are hosted on 'fast-flux' botnets where the IP address which is contacted changes every 20 minutes, forcing take-down companies to go after the domain name rather than the host; and the domains are registered in bulk, and will only be valid for less than a day.

In this new environment, counting URLs is meaningless; counting domains (the only relevant part of the URL) only demonstrates how quickly the phishing gang believes that spam filters (and 'anti-phishing toolbars') track new registrations.[1] Counting hosts is meaningless – you're merely establishing a rather inaccurate lower bound on the size of the botnet. Counting the email sending machines gives a more accurate estimate of botnet size, but isn't telling you much about phishing per se. Count the number of different Subject lines, or the number of email body texts, and once again you're measuring the gang's perceptions of spam filter technology.

Perhaps then it would be better to count phishing gangs? We attributed two-thirds of all phishing activity to just one gang [4], yet when you look more closely it's far less clear that there's actually one gang – perhaps there's several gangs who are merely using the same infrastructure, and copying each other's successful techniques?

It gets even more complicated when looking at the other third of phishing activity (measured by email sent) which accounts for 95% or more of compromised websites, free hosting accounts, and most of the other traditional measures. Since this activity is widely believed to be performed by individuals, perhaps to supplement meagre incomes in Eastern Europe, then counting each of these criminal entrepreneurs as a 'gang' skews that measure as well.

A completely different approach would be to measure how much money is being stolen. The problem now is what is meant by 'stolen'. The figures published by the UK banks are for actual losses. If money is moved out of accounts, but the fraudulent transfer detected quickly enough to undo it, then this does not count towards the statistics. Insiders say that the phishing fraud figures would double if recovery wasn't so effective. They go on to point out that the bank should actually be measuring 'funds at risk' – how much might be stolen if the money transfer

[1] One might think that new domains are only used once the old ones are suspended, but even when lifetimes are unusually long, new domains continue to be registered [1].

controls were to fail; and that anyway they've got no real idea how much fraud can be ascribed to phishing and how much to keyloggers.

Since there's no single good way of measuring phishing, it's perhaps unsurprising to see different actors choosing measurements that emphasise their importance or their effectiveness. The banks, as has just been noted, report on the smallest possible monetary metric, whilst providing a large (and ever growing) value, with no published methodology, for the number of 'attacks' they are dealing with. The security companies also publish ever growing counts of domains and URLs – to show how important their products must be.

The 'take-down' companies also count URLs even, in some cases, creatively inflating their figures. For example, sometimes all of the URLs of the form `www.example.com/~user/phishing.html` will lead to the same page – for all of the hundreds of '`example.com`' domains hosted on the same physical machine. So although just one website has been compromised, and just one URL is appearing in emails, hundreds of different attack reports will be counted. Another example occurs on the fast-flux networks where it is common for multiple banks to be attacked in parallel. The criminals are not very tidy, and so they leave the webpages for old attacks lying around. Although they may not actually be sending out further attack emails for a particular bank, the domains they're using to attack someone else are valid URLs for their old target – these URLs, which exist but do not harm anyone, are invariably included in the statistics.

The best metric, in our view, of phishing activity would be to measure the number of emails delivered into inboxes and subsequently read by individuals. Unless the individual gets to see the email, they will not visit the fraudulent website, and they cannot divulge their credentials. Unfortunately, this email data is not available – nor are the large email providers likely to publish it, since it would reflect badly on their spam filtering capabilities.

3 Bias when measuring phishing website takedown

In the first academic paper we wrote about phishing [1] (which won an APWG Best Paper award) we measured the take-down times of phishing websites using lists of URLs from `phishtank.com` and one of the take-down companies. In particular, we showed that the fast-flux hosted websites had much longer lifetimes (mean 94 hours, median 55 hours) than the other types of website (mean 58 hours, media 20 hours). People from the take-down companies were impressed by our work, and another company gave us a free 'feed' of their data – but they were less sure about the numbers we had, which they felt were very much on the high side. At the time, we suggested that they weren't taking account of the very long-tailed distribution we'd found.

The following year, we'd decided to compare the coverage of the two take-down feeds we were receiving – and discovered why our measurements differed so much from the industry perception. The take-down companies do not share their URLs with their competitors, choosing to compete on their 'coverage'. They also, fairly naturally, only remove websites for their customers, when they will be paid for their efforts.

We measured website lifetimes where the company knew the URLs and it was their customer. These had a mean lifetime of 13 hours, with a median of zero because over half the websites were removed before we could assess them. However, where the relevant take-down company was unaware of the website, but we knew the URL because their competitor had detected it, then the lifetime mean was 112 hours (median 20 hours). This of course accounted for the disparity the previous year – the companies removed the websites very fast, but only when they knew that they existed! The only remaining puzzle is why the 'unknown' websites are removed at all, and we currently believe this to be the result of reports from concerned individuals.

It might be noted in passing, that when we published this new paper [3] (which did not win a prize, even though it corrected the earlier misconceptions) we recommended that the take-down companies should share data with each other. The companies have chosen instead to emphasise the value of the banks purchasing a service from every one of them.

4 Conclusions

Measuring any sort of crime is extremely complex. If a fleeing joyrider crashes into multiple parked cars before being apprehended is that one crime or many? If a burglar steals from every house on the street is that one crime or many? If a phishing gang host their webpages on a thousand fraudulent domains, using fifty stolen credit cards to purchase them from a dozen registrars, and then transfer money out of a hundred customer accounts leading to a monetary loss in six cases: is that a 1000 crimes, or 50, or 12, or 100 or 6 ?

How we count crime and criminal activity informs our policy responses, but it's important to understand that we're prejudging some of those responses by deciding what we're going to measure – because that's going to be used to judge whether our response is being effective.

Finally, our cautionary tale of bias in phishing website lifetime measurements needs to be carefully understood. If we base policy decisions on incomplete data, or data with unacknowledged dependencies, then we will continue not to understand what is really going on, but we will be acting as if we did.

References

[1] T. Moore and R. Clayton: Examining the impact of website take-down on phishing. Second APWG eCrime Researchers Summit, Oct 2007.

[2] T. Moore and R. Clayton: Evaluating the Wisdom of Crowds in Assessing Phishing Websites. 12th International Financial Cryptography and Data Security Conference (FC08), Jan 2008.

[3] T. Moore and R. Clayton: The Consequence of Non-Cooperation in the Fight Against Phishing. Third APWG eCrime Researchers Summit, Oct 2008.

[4] T. Moore, R. Clayton and H. Stern: Temporal Correlations between Spam and Phishing Websites. 2nd USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET09). Apr 2009.