FAST, ACCURATE, COST-EFFECTIVE: CHOOSE ALL THREE

Big Data and Threat Intelligence for Phishing

A Spire Research Report

Sponsored by Malcovery Security, LLC

Executive Summary

It is too easy to get caught up in the idea of "big data" simply for its own sake. Even worse, it is common to get bogged down in technical architectures regarding it. But make no mistake - there is strong value in the analytical power associated with purposeful processing of large amounts of data. Newer techniques are being vetted constantly and provide reason to consider ways to use the techniques for threat intelligence.

The security field has use cases where sharing information has been beneficial - malware samples, breach information, etc. None of these has demonstrated overwhelming change, but they are indicators of how threat intelligence can be useful. Big data analytics provide broader, more strategic ways to contribute to the growing body of information available.

In a lot of ways, phishing is a perfect place to demonstrate the value of analytics. It has been widely studied so that we understand the pieces of the campaign and the corresponding process - from creating spoofed websites to the email campaign, and finally culminating in compromise. Companies are already sharing (pieces of) information about the problem and can clearly identify a benefit from addressing it. And there is abundant information (data) available for processing to further identify phishers and take legal action.

The old joke about the car mechanic's shop where you can have any two of "speed, low-cost, and high quality" is easily adapted to the big data world with one caveat: the promise of big data may bring with it an opportunity for choosing all THREE of "fast, accurate, and cost-effective." We will explore these concepts for the rest of this white paper.

About Spire Security

Spire Security, LLC conducts market research and analysis of information security issues. Spire's objective is to help refine enterprise security strategies by determining the best way to deploy policies, people, process, and platforms in support of an enterprise security management solution.

This white paper was commissioned by Malcovery Security, LLC. All content and assertions are the independent work and opinions of Spire Security, reflecting its history of research in security audit, design, and risk management experience.

FAST, ACCURATE, COST-EFFECTIVE: CHOOSE ALL THREE

Big Data and Threat Intelligence for Phishing

Table of Contents

Big Data: Meaningless Name; Profound Impact

"The phrase "big data" is now beyond completely meaningless. For those of us who have been in the industry long enough, the mere mention of the phrase is enough to induce a big data headache — please pass the big data Advil."

-John de Goes, "Big Data is Dead. What's Next?" Venturebeat.com[1]

"...our research suggests that the scale and scope of changes that big data are bringing about are at an inflection point, set to expand greatly, as a series of technology trends accelerate and converge. We are already seeing visible changes in the economic landscape as a result of this convergence."

- McKinsey Global Institute[2]

The meaning of the phrase "Big Data" has become so ambiguous that any conversation involving its use in a sentence must be qualified with more specifics and details just to be understood. Nowadays, it seems like it is just as likely that people using the words "Big Data" are talking about some sort of IT-nerd rapper than a revolution in information technology.

Of course, "usage-creep" for popular memes is not uncommon in technology - marketers latch onto any phrase and try to make it their own. Ironically, this often happens because the initial use case was so narrow that it didn't really fit to begin with - to "require" some specific technology like Hadoop or MapReduce in order to "make" the definition was specious on its face.

Superficiality and semantics aside, what really matters is how the use of data - big, small, or Goldilocks-sized - is being leveraged for long-term, profound impact. One thing that the big data has in common across all uses is the promise of new insights driven by data and leveraging powerful analytics.

The positive impact of big data is becoming apparent even now. In its 2011 report on big data, the McKinsey Global Institute estimates that big data may quickly become worth $300 billion a year to the health care industry, will potentially increase operating margins by 60% in the retail sector, and save over $149 billion in government spending through added efficiencies.

There are a growing number of other examples and predictions about how big data will contribute to society and industry. In the technology risk management (nee information security) field, big data needs are stunningly obvious. In the face of the "low and slow" attacks that pockmark our landscape and the complexity of the control infrastructure, much of the security intelligence is circumspect and requires

[1] 'Big data' is dead. What's next? retrieved 4/2/13 (http://venturebeat.com/2013/02/22/big-data-is-dead-whats-next/)

[2] Big data: The next frontier for innovation, competition, and productivity (http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation)

more contextual information to verify its accuracy. The field has precedent for this context and information sharing.

Information Sharing Sets the Precedent

The big data trend brings with it a renewed interest in collaboration and information sharing. What has been an "every-man-for-himself" attitude in the past has turned into one of sharing with the promise of processing even more data and reaching better conclusions about security.

Key players in the infosec field already share information to gain an upper hand on attackers. For example:

- Anti-malware companies have traditionally shared malware samples and information through their AVIEN[3] network to ensure that even in competitive environments they contribute back to the safety of the Internet as a whole. Nowadays, shared (and crowdsourced) websites like VirusTotal[4] allow private individuals to get involved as well by sharing the samples submitted with participating anti-malware companies.
- Managed security monitoring services have demonstrated the value of identifying and sharing attack information for use in protecting multiple organizations as attackers execute scans and other types of opportunistic "drive-by" activities.
- Anti-spam vendors use "spamtraps" (email honeypots) that impersonate thousands of individuals in order to collect as much unwanted and malicious email as possible. Since they are not real accounts, everything that goes to them can be considered unwanted.
- Verizon's Data Breach Investigation Report (DBIR) shares anonymized information about breaches with the community to provide more insight into the craft of the attacker and breach-related information. This sharing has been replicated by a number of companies - notably Trustwave and Mandiant.

The value of these various activities has already been demonstrated. Now, it is time to operationalize the approaches to take control of the threat.

Big Data for Threat Intelligence

One of the most exciting opportunities for the use of big data in the security field involves threat intelligence - applying analytical techniques across diverse data sets to identify sources of attacks.

For years, enterprises have focused their efforts on finding and fixing the vulnerabilities (part of their "attack surface") throughout their environments, primarily because it was the one aspect in the risk equation that could be readily controlled. Thus, organizations emphasize activities such as turning services off,

[3] http://www.avien.net/
[4] htto://www.virustotal.com/

adding more restrictive access control lists, patching known vulnerabilities, and otherwise reconfiguring systems.

The challenge with a vulnerability-oriented approach is that the supply exceeds demand. There are simply too many ways that attackers can get in. So, amidst sometimes valiant efforts by security professionals to maintain a high level of protection, the odds don't work in their favor - we don't hear even anecdotally about attackers actually being thwarted by any environment's security capabilities. Not only that, but the costs associated with a vulnerability-oriented approach can be extensive and have reached a point of diminishing marginal returns for many organizations.

Big data capabilities create an opportunity to look at another part of the risk puzzle - the part involving the threat. It is crucial to understand that the capacity of the bad guy as an "intelligent adversary" to attack one or more organizations has significant impact on the likelihood that any individual organization will be attacked.

Since many organizations have been collecting data for a number of years now, the records can be useful in identifying patterns and even sources of malicious activity that are pertinent today. Collecting and aggregating this data over time also provides future opportunities for even more benefits.

Threat Intelligence Analytics

It seems like discussions around big data tend to focus on the haystack and not on the process of extracting the needle. While it is obvious that big data requires a large volume of data, it is the analytical techniques that will drive success or failure of this trend for organizations. There are a handful of key big data analytical techniques that can serve the infosec field's ability to conduct a useful threat intelligence operation. Two key techniques are:

- Network and Cluster Analysis - Showing the relationships among certain attributes of the data set can provide important insight into the nature of a threat and their impact on the Internet.
- Inferential Analysis - Making judgment calls in a world of big data can lead to useful conclusions that can be vetted out by other information or otherwise confirmed. We used to call these bits of information "clues."

As these techniques get integrated into tools and procedures, the value proposition for threat intelligence becomes real. One early example is anti-phishing.

Using Threat Intelligence to Combat Phishing

Phishing is not a new problem. Almost 20 years from its 1996 beginning as a nuisance to America Online subscribers[5], the practice of impersonating a trusted brand via email to trick naive or unsuspecting users into visiting malicious websites is still going strong. What's more, this approach, one that essentially automates

[5] http://en.wikipedia.org/wiki/Phishing

social engineering, brings with it unprecedented economies of scale. Though some folks scoff at the weak attempts at brand impersonation made by phishers, they can send out millions of emails so that even a tiny response rate is very lucrative for them. What's worse is that the phishing emails are getting more sophisticated and more successful.

Phishing provides an obvious opportunity for threat intelligence because it is a well-known problem that has been around for a number of years, yet attempts to thwart it always seem to be one step behind the attackers who have demonstrated innovation and resilience in finding new attacks. Since there is an abundance of data that has been collected throughout the years, new analytical techniques can make a significant impact.

To further illustrate the use of big data analytics to address phishing, it is worth diving deeper into the attack process itself.

The Art of the Phish

The phisher's goal is to lure individuals to malicious but realistic-looking websites in order to steal their legitimate credentials. A typical process looks like this:

- Phisher creates a fake website by registering domains and contracting with an Internet Service Provider or compromising an existing site.
- Phisher creates the fake email to be used as a "lure" to convince victims to connect to the site.
- Phisher sends email to 10s- or 100s-of-thousands of Internet users, pretending to be a trusted brand.
- Email recipients click on embedded links that go to a realistic-looking fake website and then input their real

Who are the victims?

We all are, really. Indirectly, we all suffer from the scourge of phishers, even when we know better. Perhaps the biggest effect of phishing is the erosion of trust in the process and infrastructure. We can never take for granted that those helpful notices and statements are real; we must keep our guard up. This has long-term economic consequences. In the short term, the phished users and brands have the most to lose.

Unsuspecting Internet Users

The clearest loser in the phishing world is the Internet user who falls for the scam. Since these scams are looking for targets with liquid assets, victims often lose money from bank or other financial service accounts. Observers might think, "If they are stupid enough to fall for it, it is their own fault." But blaming the victim rarely serves the community.

Phished Brands

The 400 or so brands that get phished can also be direct victims if they have regulatory requirements or a policy to credit lost funds back to a victim simply to keep their business.

Indirectly, of course, there is even more potential for harm as the loss of confidence in a brand could lead to slower growth in future revenue when unsettled customers take their business elsewhere.

credentials.

- A script on the fake website automatically collects the credentials and saves them to a database or sends them via email to a "dropbox" account owned by the attacking phisher.
- The phisher leverages the credentials (sells them or uses them directly) which are used to connect to the legitimate website and perform actions - such as transferring funds - unbeknownst to the legitimate user.

It is easy to see that there are ample opportunities for identifying and protecting against attacks.

Phishing Countermeasures

As soon as one of the billions of email users receives the first email of a new phishing campaign, a race condition is initiated. The goal of the phisher, of course, is to trick as many consumers as possible to visit an impersonated malicious website and input their credentials for some popular brand. The goal of the brand manager is to keep that from happening. This cycle occurs over and over again.

A defender must figure out how to address the economies of scale that the attacker gains from going online. It is simply too easy for phishers to "play the numbers" when they create websites, initiate campaigns, react to countermeasures, and create a new campaign. The costs of getting into the business are extremely low and attackers quickly become proficient at "moving in the shadows" of the Internet.

Brand managers employ a number of options in their countermeasures:

- Anti-spam solutions have long employed the use of hundreds or thousands of decoy email accounts for the sole purpose of identifying spam and phishing emails. These repositories are the key element to protection as they parse an email and analyze the details for malicious URLs or binaries.
- Takedowns are performed against phisher websites that pop up. More often than not, this is a whack-a-mole challenge between the phisher and defender where the numbers overwhelmingly favor the attacker.
- Authentication and source validation efforts help to reduce risk by attempting to identify anomalous connection attempts through various risk-based authentication techniques.

All of these techniques enjoy some modicum of success, but none are sufficient to address the phishing problem. This is where big data analytics are poised to help address this problem.

Following the Trail of Breadcrumbs

You know how there's the one guy in your neighborhood that is a die-hard fisherman? He gets up early in the morning and is off creating "a river runs through it" all by himself, flyfishing his way to heaven? The most prolific phishers are like that, too. They are constantly creating new email phishing "campaigns" that employ

new brands and target different customers. As they do this, they are also looking for ways to make money in the most efficient way possible. That is when they do things that provide telltale phishing clues to those that take the time to look. These can be:

- Using the same email content patterns across campaigns
- Using the same website addresses across campaigns
- Using the same web scripts across campaigns
- Using the same email addresses and dropboxes across campaigns

These elements all create opportunities for identifying a phisher, his potential to become a repeat attacker against a brand, and his true impact in the email fraud environment.

Fast, Accurate, Cost-effective - I Want All Three

It's an old truism that out of fast, high-quality, and cheap service, you can have any two. But the two-out-of-three constraint doesn't hold when considering big data for threat intelligence. The promise of the architecture and its accompanying economies of scale provide opportunities to get all three benefits.

The following discussion highlights the expectations of big data and a way to assess new opportunities for threat intelligence in phishing countermeasures.

Fast: the need for speed

What we know about phishing campaigns is that they are time sensitive - the phisher is used to a "smash-and-grab" style campaign that occurs within hours and days rather than weeks or months. One study done on the Length of a Phishing Campaign (LPC) calculated that 66% of phishing sites had an LPC of less than 24 hours[6]. Another study showed an average of 62 hours with a median of 20 hours[7]. What's more, that same study estimated that there is an average of 18 victims on the first day that reduces by over 50% to 8 by the second day.

The good news about this phishing speed is that it demonstrates the success of takedown efforts - the phishers move this quickly because they know the reactive controls are good enough to warrant it. However, like any good "race condition" the cycle continues.

As new techniques are learned on the attacker side, the defender must get even faster with takedowns. Consider the complexity of the challenge to take down a website:

- First, you have to identify the site by analyzing emails after the campaign begins. (start the clock)
- Then, you have to contact the applicable ISP for the website.

[6] An Empirical Analysis of Phishing Blacklists, Sheng, Wardman, Warner, et. al.
[7] Examining the impact of website take-down on phishing, Clayton and Moore

- Finally, you have to wait to match priorities and timeframes with the ISP in order for them to take the site down.

After that cycle is complete and the site is down it starts all over again. With some phishing architectures, a simple command change can point the "prospects" to a new site within minutes... and the cycle continues.

The benefit of big data techniques in addressing this need for speed is that they can perform analytics on mountains of data. As more and more data becomes available, learning techniques become quicker, not slower, since the level of confidence is increased. This ensures that the evidence needed to get a site taken down can be gathered more quickly.

There is another aspect of "fast" that is worth discussing - the rate at which campaigns can be started. As data analytics gets better, and more information is provided from multiple sources, the most prolific phishers can be more readily identified and dealt with. In the same way that the McColo takedown reduced the world's spam by 75% in 2008[8], removing the world's phishers can be just as beneficial.

Accurate: how many needles are in the haystack?

There is a well-known problem in statistics called the "base-rate fallacy" that addresses the statistical challenges associated with "needle in the haystack" types of activities. The base-rate problem occurs when there are many more of one actual outcome than the other. Thus, a control test that finds 99.9% of all bad activities may still be ineffective due to an overwhelming amount of false positives. Worse, these types of false positives have a tendency to cause people to ignore them (in the same way we often ignore car alarms and store-theft sensors).

It is reasonable to consider the issue of the base-rate problem, but the data associated with phishing is already well-vetted. Two separate studies found very low rates of false positives in existing "blacklist" products[9]. This makes sense considering the abundance of "spamtrap" email accounts are out collecting nothing but inappropriate email messages.

The next step in accuracy becomes tracing the campaigns back to the phishers themselves. The goal is to combine analytics across multiple phishing campaigns and look for commonalities. Though we can identify campaigns clearly, now we want to identify the campaigner.

Accuracy of the future depends upon identifying the specific phishers and putting them out of business.

[8] http://blogs.iss.net/archive/mccolo.html
[9] Evaluating the wisdom of crowds in assessing phishing websites, Moore and Clayton

Cost-Effective: economics of phishing

Any security solution should be carefully evaluated for its cost-effectiveness. Often, this means assessing the costs and deciding whether "it is worth it" to spend the money - an approach that doesn't provide much confidence.

With phishing, however, there have been a number of studies that have provided useful numbers and estimates, providing great opportunities for understanding whether some particular control measure is cost-effective.

A recent study on phishing estimated the global losses from phishing attacks at $320 million annually, "at an absolute minimum."[10] As defenders conduct their cost-benefit analyses, then, they can compare their spending on anti-phishing solutions to the amount of losses that are allocated to them.

Trustwave estimates that phishing is 0.17% of all spam[11]. That seems like a really low number until you run the numbers. There are 140 billion emails a day, of which 75% - or 105 billion - are spam. Of that 105 billion messages, a seemingly small number are phishing emails. The problem is, that very small number still amounts to over 178 million messages per day.

Another study estimated that .4% of Internet users get phished annually, based on a study they did of 500,000 users.[12]

In economic analysis, a truly rational decision incorporates the cost of taking some action and compares it to the benefits. If the benefits are higher than the costs, then the decisionmaker may elect to take the action.

Spire ViewPoint

Seriously, do we really need to care how we define big data? No, we don't need to jump through "Hadoops" to identify a value proposition. What we need to understand is that there is power and scale in analyzing large volumes of data - the ability to process millions and billions of records in pursuit of objectives.

From a threat intelligence perspective, evolving the analytical approach to phishing by incorporating newer techniques associated with big data has already been beneficial to "frequently-phished" organizations and shows great promise in accomplishing the objective of reducing the amount of phishing activity worldwide.

The phishing problem is a perfect place to leverage analytics since the campaigns happen frequently, the accompanying information has very few associated false positives, and there is great opportunity for demonstrable cost-effectiveness.

[10] Examining the impact of website takedown on phishing. Clayton and Moore
[11] Trustwave 2013 Global Security Report
[12] Evaluating a Trial Deployment of Password Re-Use for Phishing Prevention, Florencio and Herley

Malcovery's Seven Phases of a Phishing Investigation

Malcovery leverages big data techniques to combat phishing and phishers. It has developed a "seven phase process" to evaluate all aspects of a set of campaigns and identify the phishers behind them.

Malcovery's Seven Phases of a Phishing Investigation:

1. Spam Analysis - The first step is to review the actual email messages for similarities - URLs, email addresses, etc.
2. Site Analysis - Second, Malcovery reviews individual websites to pull out elements that can be catalogued and associated with specific phishers.
3. Kit Analysis - Phishers often use the same "kit" of tools as they create new campaigns. Malcovery reviews the components of the tools and use and identifies similar pieces across campaigns.
4. Cluster Analysis - As the information develops, Malcovery evaluates the bigger picture, looking for campaign similarities and tying them back to individuals.
5. Log Analysis - as more information becomes available on the site itself, Malcovery factors in log information from all available sources, including hacked servers, victim log files, etc., to build out the complete picture.
6. Search Warrant Analysis - Malcovery assists with the search warrant process as investigations about individual phishers develop, preparing affidavits and, on law enforcement request, processing results data.
7. Open Source Intelligence - All pertinent information that is publicly available is gathered to provide support for investigations and further legal action.

The Malcovery approach provides a complete cross-campaign analytical process to drive investigations to the phishers themselves. This effectively eliminates the economies of scale that usually benefit the phishers.

Contact Spire Security

To provide feedback on this white paper or contact Spire Security, LLC about other security topics, please visit our website at www.spiresecurity.com.

This white paper was commissioned by Malcovery Security, LLC. All content and assertions are the independent work and opinions of Spire Security, reflecting its history of research in security audit, design, and consulting activities.