SCTE·ISBE CABLE-TEC EXPO 2018 | ATLANTA, GA OCTOBER 22-25

2018 Fall Technical Forum SCTE·ISBE • NCTA • CABLELABS

When Security and Privacy Collide

New Approaches are Needed

A Technical Paper prepared for SCTE•ISBE by

Sandy Wilbourn
VP Engineering
Akamai
Santa Clara, CA
+1 650 381 6129
rwilbourn@akamai.com

Craig Sprosts
Senior Director, Product Management
Akamai
Santa Clara, CA
+1 650 381 6043
csprosts@akamai.com

Table of Contents

List of Figures

Introduction

Extensive publicity about gathering and use of personal data by popular online services has increased privacy concerns, especially in developed countries. This has led to consideration and passage of privacy regulations in many parts of the world. The most visible example is the European Union's General Data Protection Regulations (GDPR) which define a new regulatory framework for the management of personal data. These regulations are colliding in unexpected ways with Internet Service Providers desire to protect their subscribers from malicious activity.

As the May 2018 deadline for implementation of the GDPR regulations drew closer security researchers realized there was potential impact on the use of the *whois* database that stores data about domain name registrations. The *whois* database was widely used for security research because it contains useful information about domain name registrations like who is registering the name, their contact information (email), location and more. Since use of domain names is fundamental to activating and maintaining most security exploits, data about their heritage is useful.

The International Corporation for Assigned Names and Numbers (ICANN), the organization responsible for administering the *whois* database, has defined a temporarily specification that pares back data fields in *whois* significantly so it is compatible with GDPR. Information that's been useful for security research in the past isn't available. ICANN has convened a group to develop a long-term solution but it's not clear where it will lead.

There are other examples in the past where proposed privacy regulations had the potential to impair security research by limiting availability of data. In early 2016 the United States Federal Communications Commission began to formulate regulations that would have restricted gathering of various kinds of network data. In this case the industry and research community collaborated and advocated for revisions that would ensure privacy while allowing for capture and use of properly anonymized network data.

It's inevitable collisions between privacy and security will continue to occur. Solving the problem of diminishing data availability means security researchers have to maximize the utility of security data that remains. Security research will need to move from rigid, deterministic, and rule-based, where personal information was helpful; to behavioral, anomalies-based analysis across very large volumes of anonymized data. The future calls for overlaying multiple layers of data where no single layer produces a result.

This will require highly automated processing and machine learning. Advanced algorithms can expand coverage of activity related to known threats, and discover previously unknown attacks, without compromising precision (generating false positives). High-performance processing of real-time data can also improve agility, or how quickly threats are found. There's also the possibility of reducing research costs by extending the efforts of human experts with machines.

This paper will cover a recent example of privacy regulation impacting security research by outlining the issues that led to the *whois* problem and compliance with GDPR. It will then discuss a way forward: applying modern data processing techniques to large data sets to expand threat coverage, improve precision, and increase agility. A production machine learning system that analyzes live streamed, anonymized, DNS data gathered from DNS resolvers serving active Internet users all over the world will be described, along with the results it can generate.

Content

1. Security Research Today

Security researchers evaluate data to find anomalies, and then cross check or validate their findings against potentially many other data source to determine whether or not a threat exists. Then they characterize the threat and publish mechanisms to deter it. Researchers use a variety of data sources to conduct their work:

- Honeypots mimic systems like mail and web servers, databases, or other services that are commonly attacked, hoping they'll be perceived as legitimate targets so malware can be captured.
- Offline data sources track ownership details for Internet resources like web hosting and cloud services, or registration data for domain names.
- In some cases other network data collected from traces or scans might be used.

The ethics of security research is a sub-field in itself, and although privacy has always been a consideration, guidelines were often based on the policies and processes of the organization doing the research since formalized privacy regulations were not necessarily directly relevant. That's beginning to change with the advent of the European Union's (EU) General Data Protection Regulations (GDPR).

2. Changing Privacy Landscape

Data privacy became a visible issue many years ago in the EU, with citizens expressing serious concerns regarding the use of their personal information by online services. After several years of debate regulations were approved by the European Parliament in April 2016 and a two-year transition period for organizations to reach compliance was established, ending in May 2018. Significant fines will be levied for businesses that don't follow GDPR guidelines. Recognizing this organizations began to assess the data they were collecting through the lens of GDPR.

The International Corporation for Assigned Names and Numbers (ICANN) is the organization responsible for administering the whois database, which contains information about domain name registrations. Whois data has been an important tool for security researchers and law enforcement agencies investigating malicious activity like phishing, malware sites, botnets and many other kinds of online crime. Information about the heritage of domain names is useful to security research because domain names are so fundamental to most exploits. More on this topic in the next section of this paper.

In late 2017 ICANN publicized a memo that concluded *whois* needed to be restructured to be in compliance with GDPR.[1] In the short term they defined a temporary specification that removed data fields containing personal information[2] like contact details, leaving only basic information like organization name, state or province, and country. They also created a multi-stakeholder working group to define a long term solution to *whois* compliance with GDPR that would meet the needs of everyone who uses the data.[3] It's at best unclear how this effort will turn out, given the complexity of the issues.[4]

[1] https://www.icann.org/en/system/files/files/gdpr-memorandum-part2-18dec17-en.pdf
[2] https://www.icann.org/en/system/files/files/proposed-gtld-registration-data-temp-specs-11may18-en.pdf
[3] https://www.icann.org/news/announcement-2018-07-02-en
[4] https://www.internetgovernance.org/2018/07/03/stacking-the-deck-the-epdp-on-the-whois-temp-spec/

A solution for the *whois* situation may be found, but heightened interest in privacy makes it likely more regulations will be implemented and it's probable other kinds of security data will also become inaccessible or obscured in ways that make it less useful. This, and widespread use of encryption, will make keeping pace with the volume and sophistication of today's security threats harder. New approaches are needed.

3. Generating Security Insights

Over the long-term balancing security research needs and privacy requirements will require making better use of available data. This means moving from rigid, deterministic, rule-based security, where personal information was helpful, to behavioral anomalies-based analysis across large volumes of data. The future calls for overlaying multiple layers of data where no single layer produces a result. Machine learning and other kinds of data processing are needed to identify increasingly sophisticated threats. Advanced algorithms can find more threats with less data, without compromising precision. New techniques can expand coverage of activity related to known threats and discover previously unknown attacks. Agility - how quickly threats are found - also improves, and research costs are reduced by extending the efforts of human experts.

A security data source that's starting to get a lot of attention is DNS resolution data sourced from resolvers or various kinds of network taps. DNS resolution data has a number of useful characteristics for security research. Domain names, and the Domain Name System (DNS) authorities and resolvers that support them, are fundamental to most security exploits. The DNS is widely used by malware developers because it connects everything on the internet, from anywhere. Virtually every network and device where an exploit might be activated will have access to the DNS. Conversely, any device that emits a DNS query known to be associated with associated malware.

The DNS has scaled remarkably as the Internet has grown and there's considerable infrastructure and tools for managing domain names. This enables highly dynamic connectivity, so exploits can move and change rapidly to avoid detection or takedowns. Malware developers can use a domain generation algorithm (DGA) to create an endless supply of random names to obfuscate their exploits. They only pay a modest fee to register the small percentage of domain names they actively use to enable their exploits. The use of DGAs is explored in depth in the paper: "A Comprehensive Study of Domain Generating Malware".[5]

From a practical standpoint DNS queries also tend to be one of the first steps in enabling malware on a host to function. A DNS query sent from a device to a known malicious destination indicates the device is associated with malicious activity, it's also usually the first "signal" that's visible on a network where it can be detected remotely. Identifying activity at this stage is extremely useful as an exploit can potentially be disrupted before it does any real damage.

This agility aspect of DNS data (and the value of DNS data more broadly for security research) was discussed in a widely publicized academic paper: "*A Lustrum of Malware Network Communication: Evolution and Insights"*. [6] The paper states: "We find that a significant percentage of malware domains can be seen in passive DNS several weeks, in many cases even months, before the actual malware sample was dynamically analyzed by the security community."

For completeness, malware developers have alternatives to the DNS. It's possible to code static IP addresses into exploits but once the address is discovered it's easy to block or takedown. Proprietary

[5] https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/plohmann
[6] https://www.computer.org/csdl/proceedings/sp/2017/5533/00/07958610.pdf

protocols can also be created to facilitate communications and management, but it requires effort to implement and maintain them, and once they're discovered they can be blocked. DNS remains the only viable choice for simple, dynamic connectivity.

4. Gathering Network Data

A data science team at Akamai began processing DNS resolution data several years ago to detect and track malicious activity. The data is from diverse worldwide sources and live streamed 24x7. It's transported over a redundant network to multiple data centers that contain parallel, intelligent processing systems the team has developed so incoming data can be evaluated in near real time.

Obtaining data is always challenging because it involves extra effort on the part of the contributors. Fortunately most resolvers can be equipped with facilities to capture query data and ship it off to other systems. There's usually a cost in terms of query performance, but it can be modest with an efficient implementation for copying query data and sending it off the server. Service providers supplying data also need to provision links to transport the data to the Akamai data centers where it will be processed as well.

5. Protecting Privacy

User privacy has always been a consideration, even in the absence of regulations. Another advantage of using DNS queries gathered from resolvers as a security data source is it's minimally invasive of privacy. Unlike technologies that promiscuously gather and evaluate traffic in the data plane, DNS queries only contain source/destination IP addresses and domain name related data. Personally Identifiable Information (PII) like IP addresses can be anonymized so that it cannot be traced to an individual. This topic will be addressed in the next section.

Data used for research at Akamai is anonymized with the Lucent extension to Crypto-PAn, a well-known cryptography-based sanitization tool for anonymizing IP addresses. Service providers who own the resolvers control all aspects of the anonymization of their query data; they configure which potential PII is anonymized and create and manage the anonymization keys. A third party cannot reverse the anonymization, only the provider can, using keys they generate. They use one key for anonymizing all of the data in their network which is a bare passphrase consisting of any ASCII text, on any number of lines.

Data is also encrypted in transit. This requires provider systems to initiate secure connections to the destination servers. OpenSSL has proven to be a good solution. Connections from the provider network to Akamai servers are authenticated by looking up a host specified in an authtoken file supplied by Akamai, connecting to the host using TLS, and exchanging and verifying certificates.

6. Building a Layered, Intelligent Processing System

The team set goals of improving threat coverage and precision by applying intelligent processing to the DNS data. Another objective was to do all of the processing in near real time, so threats could be identified, validated, and published as quickly as possible. This has led to development of a number of systems for intelligent processing, summarized below and described in detail in the following sections.

- Preprocess the data to reduce noise so more processing power can be applied to data of interest
- Map relationships between domain names
- Assign domain reputation scores by joining with other data sources to evaluate "maliciousness"
- Expand coverage using techniques similar to natural language processing
- Correlate relationships between malicious domain names

- Visualize clusters using 2D and 3D graphs to better understand relationships

Each system operates as a separate "layer", with each adding intelligence to the findings of others. In most cases no single layer offers conclusive evidence that a domain name is malicious, instead they all work together to formulate conclusions. In effect the network, or more accurately streamed network data, looks like a massive, extremely diverse, near real-time honeypot.

7. Discovering New Core Domains

One of the earliest revelations of the research was newly observed domain names tend to be more highly correlated with malicious activity. This makes sense intuitively because malware developers need to constantly change the face of their exploits to avoid detection and take down. One of the ways they do this is to constantly change the domain names associated with their exploits.
In creating a methodical approach for studying newly observed domain names we defined the concept of a "core" domain, which is also known as an "effective 2nd level domain" (e2LD). For instance: *www.example1.com* and *www.example2.co.uk* are core domains or e2LDs. It can be seen that core domains usually capture domain ownership. For the past 5 years Akamai researchers have been tracking new core domains, essentially newly observed domain names, and in 2017 undertook a project to greatly improve the infrastructure in order to study them more intensively. Details of the new core domain work were presented at a DNS conference in 2017. [7]

The team developed a read/write in-memory processing engine that was capable of operating on a 1.5 million QPS data stream (scalable as the data stream grows). This engine was designed to enable real time processing to reduce noise in the data and evaluate the relevance of each query. Algorithms also flag other kinds of anomalous behavior, such as incoming queries for domains with query patterns that substantially differ from previous patterns. This engine effectively detects potential phishing, bot and other malware activity, DNS based DDoS attacks, and DNS tunnels.

[7] https://indico.dns-oarc.net/event/27/contributions/456/

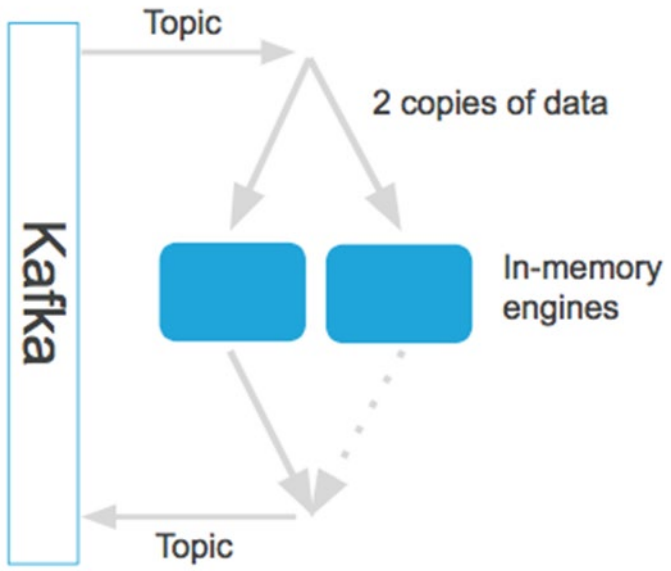

Figure 1 - A read/write in-memory processing engine

A read/write in-memory processing engine processes live streamed DNS queries at 1.5 million queries per second. This engine reduces noise in the data and evaluates the relevance of each query.

Output of the processing can be seen in Figure 2 below. A dashboard displays a number of statistics such as total queries processed, new core domains found, queries to new core domains that resolve and don't resolve (return an error code). The resolution status is represented in dark blue (resolved) vs light-blue (not resolved). Seeing the resolution status, including the answer itself, is useful because it turns out domains that are not registered are frequently used by botnets through DGA (domain generation algorithms). While blocking such domains may or may not help, there is still value in identifying an infected machine even if a query did not resolve.

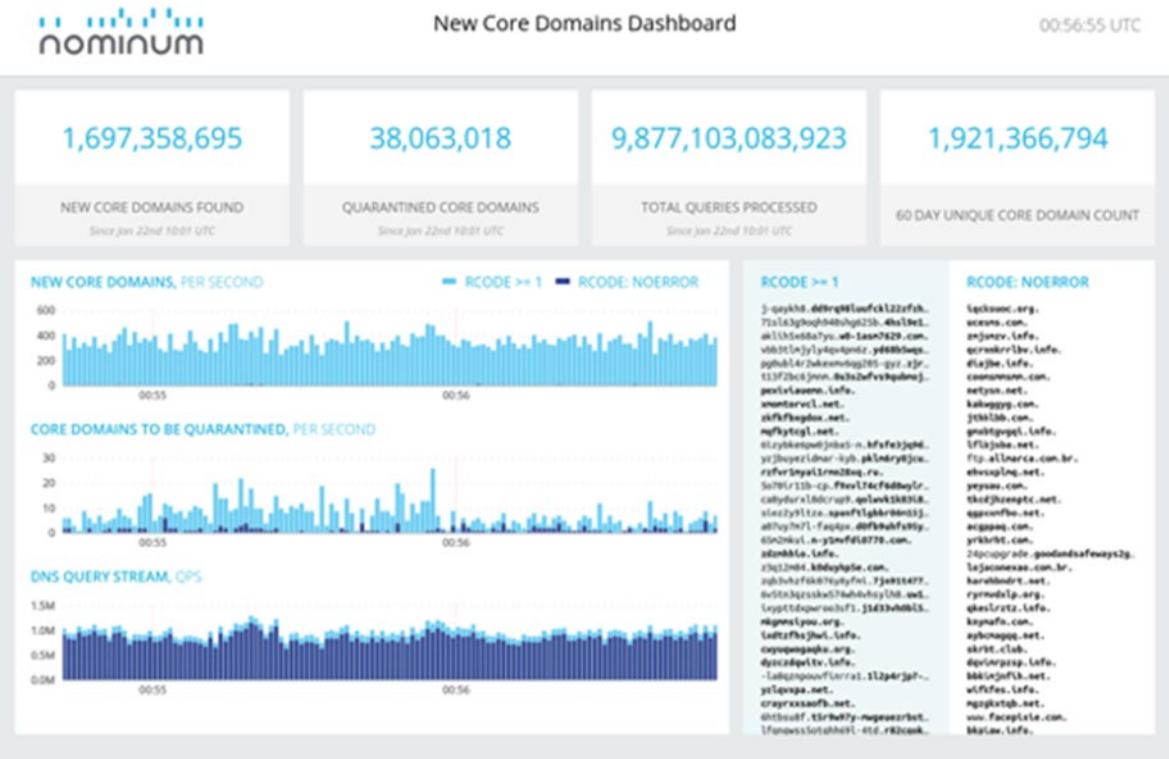

Figure 2 - Output of a near real time processing engine evaluating live streamed DNS resolution traffic

Output of a near real time processing engine evaluating live streamed DNS resolution traffic to find newly observed domain names. Domains discovered by this engine are subject to additional processing to validate their maliciousness and characterize their behavior.

Domains that will be quarantined are also displayed. These are the targets for more processing to determine whether or not they are actually malicious. Subsequent processing also reveals their intent so they can be categorized.

New core domains provide useful information about other vectors of attacks. For example, an uptick in the use of social media for distributing malware was uncovered. In this case waves of free airline ticket promotions by airlines. The domain names for the promotion used alternate character sets that looked like ordinary roman characters in order to trick users. The domain name in the url displayed a subtly different character but the actual domain name seen by the resolver was much different and easily detected by the new core domain logic.

8. Adapting Natural Language Processing to Domain Names

Security list providers catch some of the domain names an exploit uses from honeypots, but they typically don't capture all of the names in use. Malware can also include anti-honeypot techniques to fool the honeypot, for instance `pykspa` uses real and fake DGAs to confuse the honeypot output.
To expand coverage of malicious activity generated from the other layers in the system additional techniques borrowed from natural language processing are used to reveal relationships among seemingly random domain names and clients that query them. The model borrows concepts from the word2vec work

done at Google. [8] Quarantined domain streams are fed into the model and it generates clusters which group the most correlated names together. The model applies an advanced neural network structure onto the original word2vec neural network by modeling the DNS query sequence and discovering the in-depth correlation among domain names in a massive DNS traffic stream.

9. Evaluating Quarantined Domains

Clusters that are discovered are validated using 3rd party security lists, typically generated by human researchers, which include malware C&C, malvertising, phishing, etc. Relationships between domains are mapped (analogous to a social graph) so likely neighboring domains that are malicious can be propagated. Algorithms overseeing these layers of guilt by association generate a Domain Reputation Score that categorizes domains to be designated as malicious or those worthy of even more analysis.

Measurements calculated by the research team showed propagating human security intelligence to clusters discovered with machine learning can expand coverage by 5x to 10x. To the point made earlier about the agility of DNS resolution data, malicious clusters are also regularly identified which didn't appear until hours or days later on 3rd party threat lists.

Malicious or suspicious domain names discovered in data Akamai collects are stored in a reputation knowledge-base. Continuous improvements to this database make associated machine-learning systems faster and more accurate so more malware can be effectively blocked before it causes damage.

Looking at the output of the algorithms for individual threats and then doing a deeper dive to see what other kinds of patterns emerge always offers interesting insight. For example, additional analysis of the machines that emitted the kill-switch domain for Wannacry showed there was a significant correlation with gaming use of those machines and TeamViewer, a tool for remote administration. This makes sense since leaving certain ports open increases the likelihood exploits will get into unpatched systems.

As another example, an evaluation of Petya's time sequence showed it took the dropper exactly 2 minutes from the time it was downloaded until it started querying the payload site. Only a couple of minutes! AV for these infected users, assuming there was one installed, didn't catch the dropper file. Instead it allowed it to install itself, and then make a query to the payload site.

10. Visualizing Security Data

Continuous improvements in graphing technology allow better visualization of threat activity. Results calculated using the correlation techniques above are fed into a model that groups the most correlated domain names together into clusters and places them on special 2D and 3D graphs so their relationships can be better understood. An example of the graphs that can be generated are shown in the figure below.

[8] https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

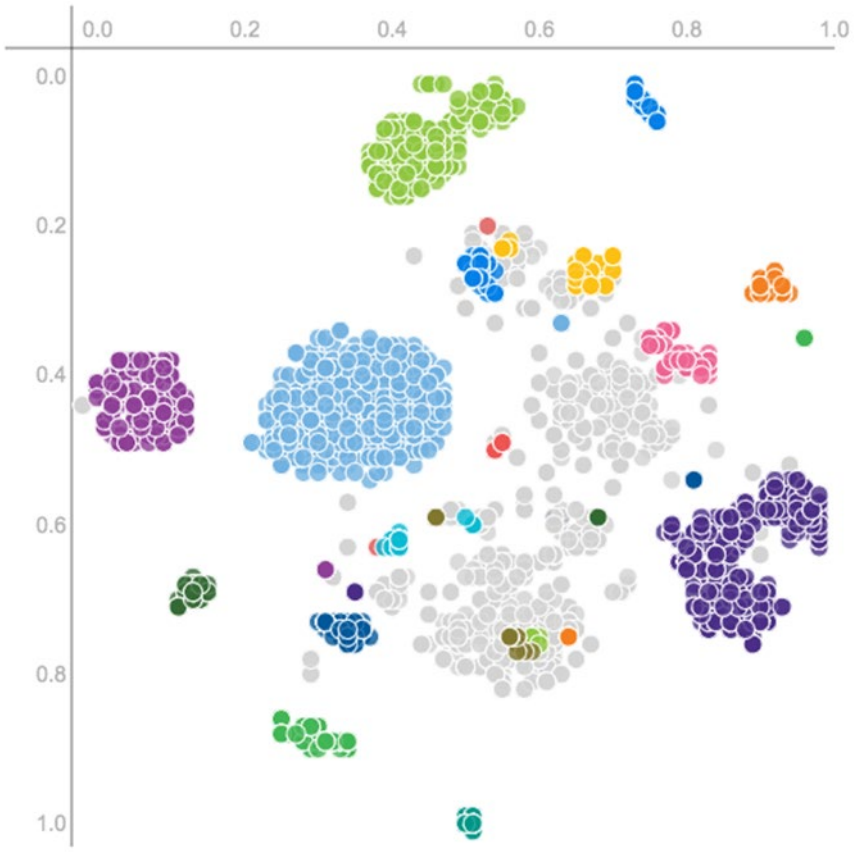

Figure 3 - Two-dimensional visualization of results from a clustering engine

Two-dimensional visualization of results from a clustering engine that uses unsupervised machine learning to correlate characteristics that reflect association with a common underlying threat.

Conclusion

It's inevitable collisions between privacy and security will continue to occur. This will make maintaining security in a privacy driven world harder. The telecommunications industry needs to monitor regulatory initiatives worldwide and advocate for policies that preserve privacy but allow for gathering and use of data used for security research. Even with advocacy, security data sources will continue to be obscured or blocked altogether, and with stiff penalties for privacy violations creators of data are likely to become more cautious so raw data will be less available, more opaque, and generally less useful. Yet malware developers won't relent, in some cases they've already implemented with multi-faceted exploits that evade defenses and propagate rapidly. Maintaining an edge will require making the most of available data sources.

Almost all threats have a footprint in the DNS and analysis of query traffic captured by DNS resolvers can provide early detection of malware since exploits have to resolve addresses of malicious resources under their control before they can use the functions they rely on to operate and propagate. Evaluating DNS query data also offers the possibility of improving coverage of diverse communications channels malware uses. Human driven security research will always be necessary but intelligent processing and machine learning will become essential tools that strongly complement agile, rich and diverse DNS data.