White Paper

EMERGING THREATS

The Rise of Dynamic Malware

Using Data Science to Prevent Cyberthreats

Introduction

As the number of Internet-connected devices continues to explode, malware threats are growing. Malicious code authors are becoming more sophisticated to evade detection, employing techniques such as adaptive computer code and changing command/control server locations. Fresh approaches to online consumer protection are required to tackle these increasing cyberthreats.

The Internet of Things (IoT) is bringing cyberthreats to more devices and, potentially, home information. As attackers adapt malware to IoT, it will introduce new challenges to consumers and service providers. The primary focus of IoT to date has been on innovation, not security. “No one wants to build security into their devices, because no one is going to pay more for a secure device,” said Bo Rotoni, Co-Director of the Institute for Information Security and Privacy at Georgia Tech University.[1]

Limitations of traditional security approaches detecting “zero-day” malware threats—those threats that arrive before developers have time to release a patch—are well documented. For example, one report showed anti-virus software only detected 51 percent of zero-day malware samples as threats.[2] These evolving web-based threats leave consumers vulnerable. Kaspersky Labs reported widespread presence of the Locky ransomware virus in Q1 2016, with infections in 114 countries[3] —while the virus continues to spread. Locky is known for changing continuously, making it difficult to deter.

Although malware developers go to great lengths to obscure their exploits, most rely on the Domain Name System (DNS) because it is readily available in every service provider network. The presence of malicious DNS queries thus signals the presence of malicious activity. Nominum Data Science has a unique vantage point

Nominum Data Science is a worldwide team with expertise in Internet security, machine learning, artificial intelligence, natural language processing and neural networks. The team analyzes more than 100 billion anonymized DNS queries every day to avert cyberthreats.

[1] http://www.iisp.gatech.edu/sites/default/files/documents/2016_georgiatech_cyberthreatsreport_onlinescroll.pdf

[2] Lastline Labs, 2014 http://labs.lastline.com/lastline-labs-av-isnt-dead-it-just-cant-keep-up

[3] Kaspersky Labs SecureList IT threat evolution in Q1 2016 https://securelist.com/analysis/quarterly-malware-reports/74640/it-threat-evolution-in-q1-2016/

as it processes more than 100 billion DNS queries daily, in real time, using data shared by customers over our global network. This enables Nominum to identify attacks that are not published elsewhere or before they are made public through other research.

Malware and DNS

Computers and other devices become infected with malicious software in a variety of ways. Social engineering, or tricking users into activating an exploit, is increasingly prevalent. Infected devices are used for different purposes—many of which have a visible impact on consumers—such as locking up computer files until a ransom is paid, stealing personal information and selling it or using it to gain access to financial assets. Malware can also be used to attack provider systems directly (DDoS) or to send spam. Malware inside provider networks is typically controlled remotely and goes from dormant to active instantly and unpredictably. Rapid response is critical.

The DNS is a distributed naming system for computers, services, or any resource connected to the Internet or private network. It translates human-friendly domain names (e.g. www.nominum.com) to numeric IP addresses used to route network traffic. DNS is used in all facets of the Internet such as web browsing, setting up cellular and VoIP calls, routing email, retrieving application updates and connecting/managing devices on IoT.

As a highly robust, reliable and ubiquitous Internet protocol, DNS provides criminals and profiteers with a ubiquitous platform to launch and manage a wide range of exploits. DNS data analysis can reveal these malicious activities—including the infamous DNS tunneling and DNS-based DDoS attacks—to serve as an efficient and effective "early warning system" for identifying malicious network activity in near real time.

Nominum Data Science research from mid-2015 shows interesting trends that are covered in this paper. The biggest change Nominum has seen is the increasingly agile nature of malware. Highly dynamic threats on the Internet have been identified from DNS data as evidenced in this paper.

Nominum Data Science processes more than 100 billion DNS queries daily in real time

The Increasingly Dynamic Nature of Attacks

Continuous innovation by cybercriminals can be seen in the one-week sample of DNS query data from late 2015 shown in Chart 1. An average of 73,000 new core domains are observed each hour. The number of domains varies widely, with nearly 180,000 new domains observed during some hours. New domain names are highly correlated with malicious activity because attackers need to constantly register new names as older names are discovered and used to block their exploits.

Obvious questions arise from this data: What are these domains used for? How effective are existing approaches to detecting malicious domains?

To answer these questions, Nominum analyzed a single 24-hour period of DNS data from the week where we saw:

- 1,005,887 unique domains queried
- 858,174 of these domains had only one query made against them
- 2,670 domains reported as malicious by a sample of leading anti-malware vendors at the time (a fraction of the total)

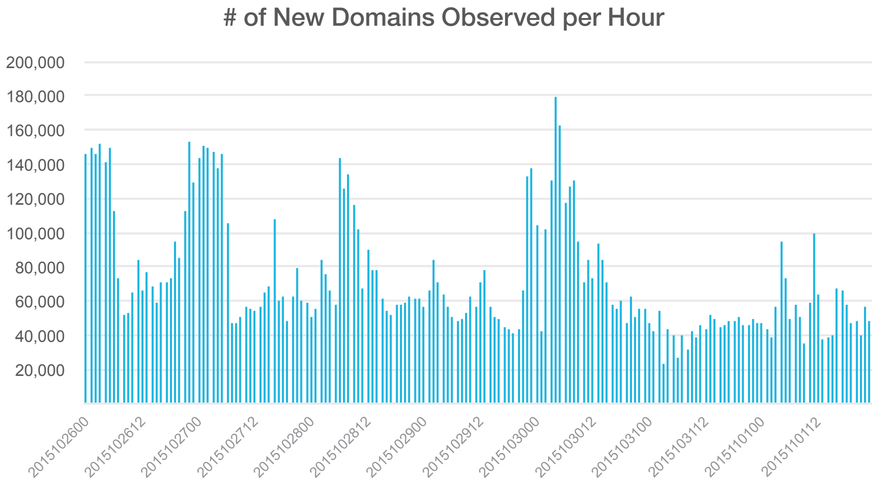

Chart 1: New domain names strongly correlate with malicious activity. Hundreds of thousands of new names may appear each day and are candidates for further analysis.

Drilling into the data, there is clear evidence of anomalous activity. A simple but useful technique for identifying anomalies in DNS data is to look for queries to domains generated by Domain Generation Algorithms (DGAs). These programs automatically generate domain names used in attacks. One way to identify DGAs is to evaluate the length of domain names. Chart 2 plots the number of domains seen in a given day grouped by the length of the core domain name.

Comparing this data set against malicious domains tabulated on existing threat lists shows only a small fraction (0.27 percent) of the domains match. What about the rest?

Searching for Anomalies

A couple of anomalies stand out: the number of domains with 14 characters (the large spike on the chart), 17 characters and 37 characters exceed what would be expected in a typical distribution. More precise determination of maliciousness requires applying machine learning techniques to a larger set of features. Drilling into the data, there are 14 character names like oiceadybpb.net, ehslrsbwby.com, fgirfpwrr.info and 37 character names like ac94c74043d0d15b51e8ef970cc56540aa.toh and b2ab08e9f7349c62bcca7882af4e799953.in. It's obvious these names weren't designed to be understood by humans. But that alone is insufficient evidence that a name is malicious. Applications often use names that aren't human-friendly

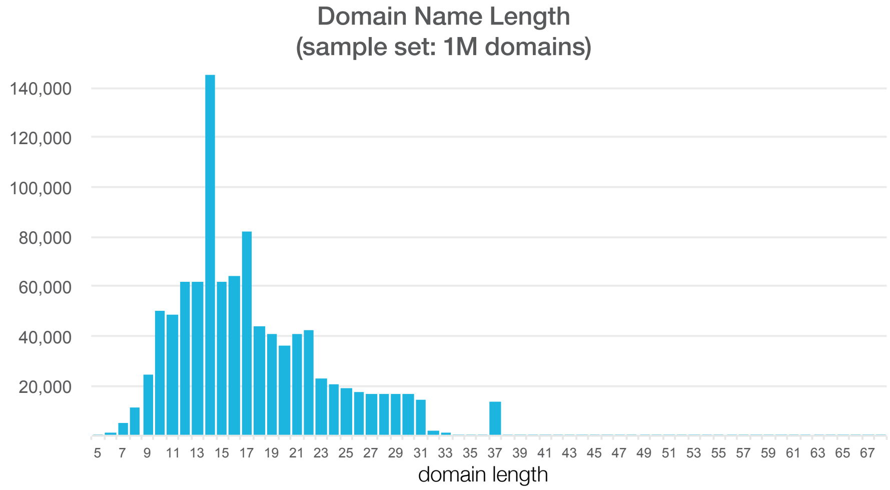

Chart 2: Evaluating the length of domain names is one way to view anomalous activity. This chart shows an unusual spikes. These names are flagged for further analysis.

for updates, or to point to files.

A sample set of characteristics to be further analyzed include:

- domain hosting location
- which clients query the domain
- co-occurrence of certain domains with other domains
- information about when the domain is queried, and changes in query volume

When machine learning is applied to the same set of DNS queries, the results on a two-dimensional graph show "clusters" of malicious activity as illustrated in Chart 3.

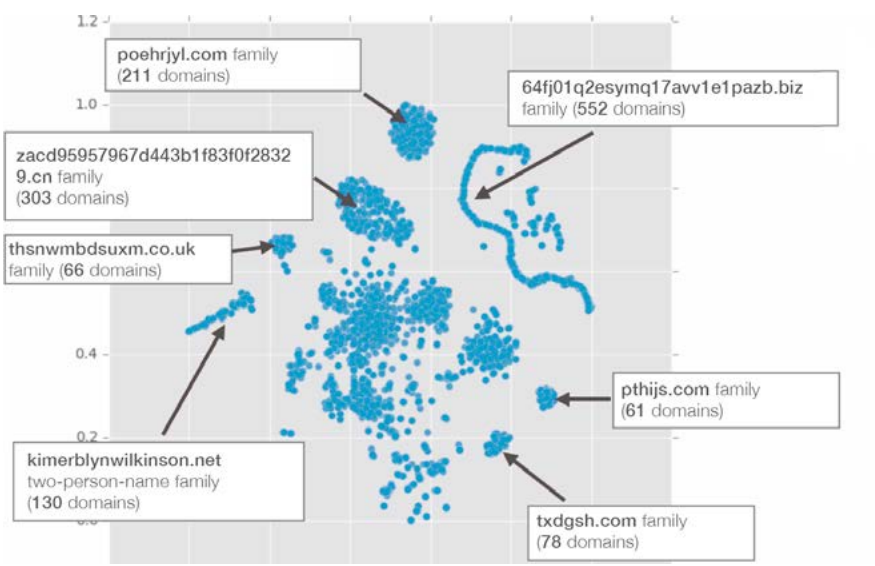

Chart 3: Correlation technology identifies names with common characteristics. Using this technique, entire families of exploits can be revealed.

While the specific machine learning techniques used to generate the clusters are beyond the scope of this paper, they represent the output of a multidimensional feature matrix into a two-dimensional space. Each dot represents a domain, and the closer the dots are to each other, the more similar they are. Seven different clusters are highlighted, each representing unique sets of malicious domains, likely under control of the same criminal actor.

To better understand the nature of the clusters, Nominum tracked queries to domains on two of these clusters over time using a metric called "Nom-Rank." Nom-Rank measures the popularity of a domain based on several factors including the frequency of queries to that domain and number of clients querying it. Chart 4 tracks the Nom-Rank score of malicious domains in two clusters over time.

Domains in these clusters are used for less than three days after the day they are first queried. Since domains are no longer actively used in exploits, any names that appear on a threat list after that time are useless in protecting against attacks. Malware developers have already activated new domains. This also illustrates why "newness" is such a strong correlating factor for maliciousness.

DNS-based machine learning techniques employed by Nominum Data Science on this data set identified 1,788 domains on October 31 as malicious; of these, only 269 names had been detected by third party anti-malware vendors as of November 1st. Dynamic DNS-based detection found additional malicious domains that would

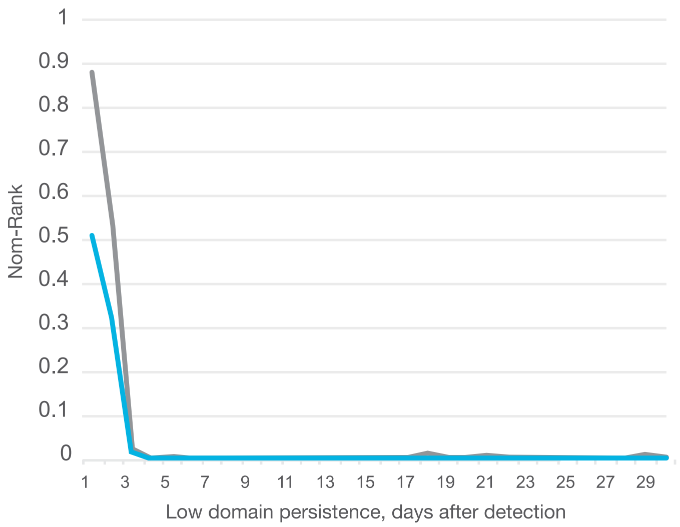

Chart 4: Query traffic to two sample clusters of malicious domains shows they are only active for a few days.

otherwise be undetected, or detected too late.

What about the more than one million other domains seen in the data set? They were generated with malicious intent but never activated. In fact, this is another way attackers obscure their exploits. There is considerable “noise” researchers need to work around in order to see the “signal.” Nominum Data Science continues to see malicious activity related to these same clusters and experience suggests the number of activated malicious domains will increase significantly in the coming months. Exploits don’t go away, they just change continuously, adapting to the shifting security landscape around them.

Summary

Cybercriminals with substantial technical expertise continue to innovate so their exploits remain viable and profitable. A key part of their success depends on rapidly changing domains used to control their infrastructure, and hiding small numbers of active malicious domains amongst millions of new domains. These techniques make it difficult for traditional security solutions to detect such threats.

This report shows how DNS-based detection identifies malicious activity quickly, in seconds rather than days. It also shows that DNS-based clustering algorithms, seeded with a single domain known to be malicious through traditional approaches, can enable discovery of thousands of malicious domains that would otherwise remain undetected.

DNS-based protections closely track underlying threat dynamics and keep subscribers’ homes and devices safer from threats like ransomware, or theft of valuable personal information. Dynamic attacks targeting homes filled with connected devices make the need for dynamic protections more urgent.

About Nominum

Nominum is the world's DNS innovation leader and the first company to create an integrated suite of DNS-based, subscriber-centric applications to digitally transform service providers and personalize the online experience.

Nominum N2™ solutions leverage the company's market-leading Vantio™ DNS software and expert team of data scientists to forge a clear path for service providers to move beyond a one-size-fits-all, network-centric approach to a value proposition that is highly differentiated and subscriber-centric. N2 provides an extensible framework that synchronizes digital capabilities with people, processes and systems across the organization, delivering personalized solutions that enhance subscriber value and brand loyalty, fuel revenue growth and bolster competitive advantage.

Nominum is a global software company headquartered in Redwood City, California. More than 100 service providers in over 40 countries trust Nominum to enable a safer, more personalized Internet experience and promote greater value to subscribers. Nominum DNS software resolves 1.6 trillion queries around the globe each day—roughly 100 times more transactions than the combined daily volume of tweets, likes, and searches taking place online. For more information, please visit nominum.com.

About Nominum Data Science

Nominum Data Science is a worldwide team with experts in Internet security, machine learning, artificial intelligence, natural language processing and neural networks. Previous projects of team members include quantum physics and data analytics used to discover the Higgs boson at CERN and some of the earliest investigations into the structure and propagation of botnets. Examples of threats tracked by Nominum Data Science include:

- DNS-based DDoS attacks that use millions of home gateways with open DNS proxies, or bots in infected devices that compromise networks
- Sophisticated malware and bots, secretly loaded onto consumer devices, that rely on DNS to trigger spam, financial and personal theft, and more
- DNS tunnels carrying other protocols that use special client software to steal service

For more information about Nominum Data Science and cyberthreat protection, please visit nominum.com.

© 2016 Nominum, Inc. Nominum, Vantio and N2 are trademarks of Nominum, Inc.

CORPORATE HEADQUARTERS

Nominum, Inc.
800 Bridge Parkway, Suite 100
Redwood City, CA 94065

+1 (650) 381-2000

hello@nominum.com