All-packet MONitor (AMON)
Internet Anomaly Detection in Multi-gigabit Streams

Victor Ongkowijaya (U-M CSE), Owen Webb (U-M CS-LSA), Michael Kallitsis (U-M, Merit Network Inc.)

Abstract

The Internet is a system of interconnected networks involving an immense and continuous flow of information. With the rapid proliferation of Internet-connected devices, it is crucial for network security professionals to have the necessary tools and software to assess cyber-risk, and protect their networks and systems. This study aims to extend All-packet Monitor (AMON), an open-source framework for online monitoring and analysis of multi-gigabit network streams developed by researchers at University of Michigan and Merit Network (which operates Michigan's research and education network). AMON allows operators to quickly visualize and diagnose attacks, and a prototype has been already deployed at Merit Network, currently processing 10Gbps+ of live Internet traffic. In this study we are working on building new AMON monitoring modules that could allow us to process DNS (Domain Name System) traffic, aiming to identify malicious activity such as malware propagation within a network. In particular, we are working on efficient visualization tools that would help operators quickly identify the onset of such activities.

Introduction & Background

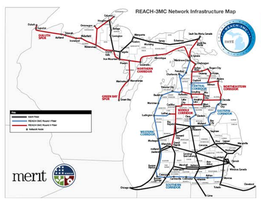


Figure 1: Merit backbone

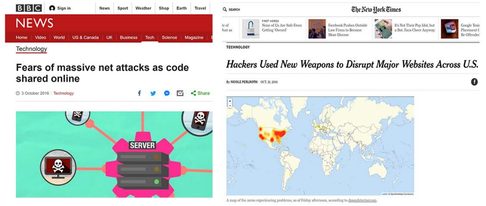


Figure 2 & 3: Recent network attacks

Merit is a academic internet service provider (ISP) that serves schools and other non-profit organizations across Michigan. This study focuses on a specific type of internet traffic data sourced from Merit, namely, Domain Name System (DNS) traffic. DNS is a vital component of internet infrastructure because it acts as the internet's 'address book' allowing computers to translate between human-understandable website domains like "www.google.com" and computer-understandable IP Addresses like "10.2.8.8." DNS' importance makes it a prime target for attackers utilizing Denial of Service attacks and botnets. One such attack recently occurred in October of 2016 when a malware named Mirai was used to attack Dyn DNS servers bringing down many popular websites like Spotify, Github, and Paypal for hours.

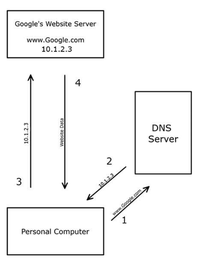


Figure 4: DNS

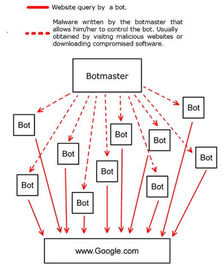


Figure 5: Botnet attack structure

The Domain Name System (left)

*1.*When you request a website, your computer first asks a DNS server for the IP address of the website. Similar to asking for directions.

*2.*The DNS server responds with the correct IP address.

*3.*Your computer uses the IP address to request the website data from the proper server.

*4.*The server, Google's for instance, responds with the website you requested.

Distributed Denial of Service Attack (right)

Botmaster: A malicious attacker, usually a singular person.

Bot: An infected computer in the broader botnet, controlled by the botmaster. Could be any device with an internet connection.

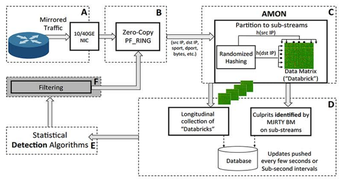
Website (www.Google.com): Website servers can be taken down if they receive too many requests at once. Also known as a Distributed Denial of Service (DDoS) attack.

Objectives

Develop an open source platform that will be broadly available to network operators, and is based on inexpensive hardware. This is because commercially available tools can be prohibitively expensive. It will also:

- Focus on real time, raw packet data.
- Detect malicious activity, such as botnets, which is the particular aim of this study.

Proposed Methodology for Intrusion Detection



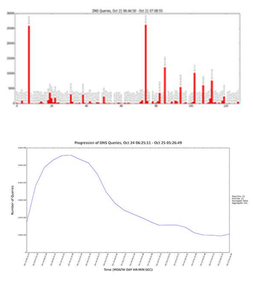


Figure 7: Data structure visualization

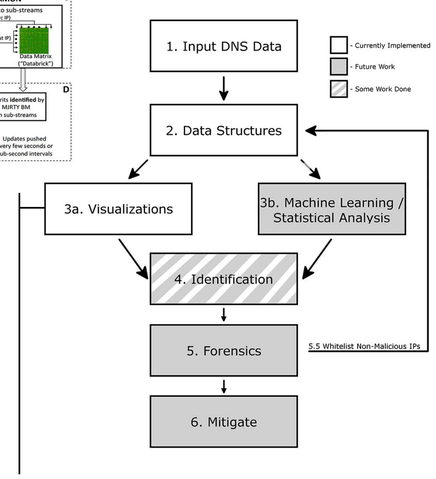


Figure 6: Intrusion detection methodology

1. Internet traffic data is collected at Merit Network, which is then accessed by researchers.
2. Data is filtered and stored in appropriate data structures. These must be computationally and memory efficient.

3a. Visualizations are generated (databricks, histogram, time-series) based on the data structures.

3b. Machine learning and statistical methods are applied to the data structures to ensure constant threat monitoring and automatic detection of malicious activity.

4. Suspect IP addresses are identified using efficient algorithms.
5. Forensic techniques are applied to classify malicious activity, and then associated with suspect IP.

5.5. If forensics classify certain activity/behavior as non-malicious, a whitelist is created for this classification/IP. This whitelist is used for future data structure generation.

6. Measures are taken to mitigate threat.

Findings

Statistics

- Approximately 2,000,000,000 queries per month.
- 66 million per day, 2.7 million per hour, 46 thousand per minute.
- Around 5000 unique IPs, which is low because each IP "represents" many IPs.
- More than 25 million unique domains queried for.

Databricks

- Parse components from data stream. Core components are source IP and target domain. As we add more functionalities, we parse for more components such as rcode and ID.
- IP and domain are hashed using md5 to buckets 0-127.
- Create a 2-dimensional array, with IP as column index and domain as row index.
- The content of this array is the frequency -- how many times did this IP query for this domain?
- The generated data structure is then visualized into a heatmap. Higher frequencies are red, lower frequencies are green.

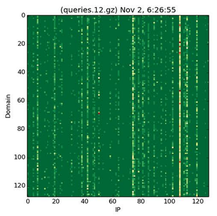


Figure 8: Normal behavior, light

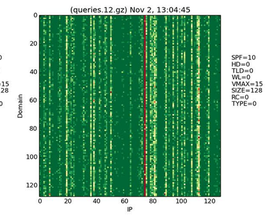


Figure 9: Normal behavior, heavy

Nature of Data

- Visualization shows that DNS traffic is noisy -- there are many vertical lines. This is initially unexpected and provides some issues with detecting anomalies and malicious attacks. The cause of these vertical lines is thought to be benign, but active, entities -- such as mail servers and DNS servers of university institutions.
- There are many strange activities. Processing real world data has many unknowns and is not clean.

Reducing Noise

- Top level drop
- Heavy hitters
- Whitelisting
- Rcode
- Query type

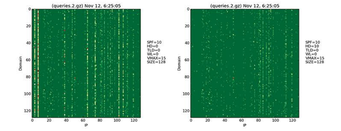


Figure 10: Cleaning top 10 heavy hitters

Detecting Anomalies

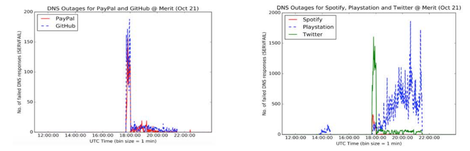


Figure 11: Mirai SERVFAIL DNS responses

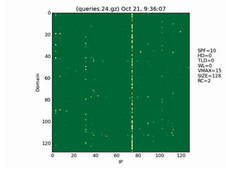


Figure 12: Mirai SERVFAIL databrick

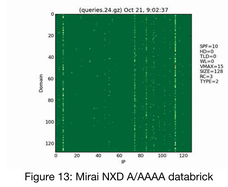


Figure 13: Mirai NXD A/AAAA databrick

Conclusions & Future Work

Products

Tools developed:
- Establish DNS collection infrastructure.
- Generate databricks given DNS traffic data, which is also able to filter based on different parameters such as query type, rcode, and number of heavy hitters.
- Generate heavy hitter histogram.
- Identify heavy hitters.
- Data characterization/exploration.

Information:
- Documentation on format of traffic data.
- Visualization and statistics on certain time frames of interest.
- Documented insight gained from analyzing data.

Conclusion

- Different data structures are needed for different data types and purposes.
- DNS data can be revealing, and we can learn many things about the network.
- Anomaly detection associated with DNS traffic is difficult. Benign heavy hitters obscure the activity of potentially malicious attackers, and the data is not clean.
- Some future related work include further studies on DNS, as well as particularly on Internet of Things devices.

References:

Han Zhang, Manaf Gharaibeh, Spiros Thanasoulas, and Christos Papadopoulos. BotDigger: Detecting DGA Bots in a Single Network. January 2016.

Kensuke Fukuda and John Heidemann. Detecting Malicious Activity with DNS Backscatter. October 2015.

Michael Kallitsis, Stilian Stoev, Shrijita Bhattacharya, and George Michailidis. AMON: An Open Source Architecture for Online Monitoring, Statistical Analysis and Forensics of Multi-gigabit Streams. January 2016.