An study of the concepts necessary to create, as well as the implementation of, a flexible data processing and reporting engine for large datasets.

Ignus van Zyl

1 Statement of problem

Network telescopes observe millions of packets of Internet traffic through the monitoring of unused IP address space Moore et al. [2004]. The data gathered by these telescopes is important to the study of Internet Background RadiationPang et al. [2004] and also more specifically to the monitoring and study of potentially malicious IP traffic across the internet Irwin [2011]. With increasingly large datasets being created by telescopes across multiple studies, the distribution of analysis results through the scientific community could be improved by creating a flexible yet standardised reporting engine Irwin [2013]. Such a standardised reporting format will also provide researchers with an overview of trends in the analysed data-sets.

2 Objective of research

To utilize data analysis techniques to process large packet capture data-sets, and to create output in the style of a report that succinctly summarises the key trends and characteristics of the data-sets. Such an engine should be flexible with the data it can process. This engine should also aim to offer customisable report formats, as well as suitable high level statistical and graphical representations of the target data-set.

3 Background information on subject material

3.1 Network Telescope

A network telescope is some part of assigned and address routed IP space. The IP addresses are used but not utilized by any host system Moore et al. [2004]. Any traffic received can be defined as unsolicited Moore et al. [2004].

Studying traffic captured by these telescopes provide researchers with interesting insights into network security events such as denial-of-service attacks, Internet worm packets and network scanning attempts Moore et al. [2004]. Network telescopes do not respond to requests, nor send traffic within the IP address block that it observes. The telescope does not record packet data outside of the observed net block Du and Yang [2011].

3.2 Internet Background Radiation

Internet Background Radiation is the formal term given to unsolicited Internet traffic that is sent to unused IP addresses [Pang et al., 2004]. This traffic can be charecterised as non-productive as the packets are destined for unused addresses and servers that are not running [Pang et al., 2004]. This traffic can consist of a number of different packets, including but not limited to: denial-of-service backscatter, network scans, worm infection attempts and corrupted packets Pang et al. [2004]. The purpose of network telescopes is to capture this radiation; which can then be subject to further analysis.

3.3 Measured Network Telescope traffic

A typical network telescope will collect both relevant and irrelevant traffic. One task of telescope data-set analysis is to sort packets so that those that offer no intelligible information can be identified and ignored. The packets worth analysing can be grouped as follows:

Active Traffic

Active traffic can be determined to be packets which are expected to elicit a response when processed by the TCP/IP stack of the receiving host [Irwin, 2011]. TCP packets that have the SYN flag set, indicating that a response is expected, are counted for analysis purposes. TCP packets that have the ACK or RST flags set could be the result of backscatter traffic. Backscatter traffic could be the result of the monitered address being used in spoof packets as part of a security attack [Irwin, 2013]. ICMP traffic packets that attempt to elicit any response will also be analysed [Harder et al., 2006, Irwin, 2011]. Since UDP traffic is stateless, no expectations for initiation or response can be inferred from the headers. As such, deeper analysis of the UDP payload is required to determine if the UDP traffic is active [Irwin, 2011].

Passive Traffic

Passive traffic is traffic that, when processed by the TCP/IP stack of the receiving host will give no legitimate response. It is considered unlikely that

potentially malicious software would use these packets to determine information about the host system Irwin [2011]. Passive traffic is usually the result of scanning activity; which is typically the result of the monitored IP range being spoofed, denial-of-service flooding, misconfiguration of network addresses or mangled and unintelligable packets [Irwin, 2011].

3.4 Important Statistics and Identifiers

The analysis will focus on identifying and quantifying the telescope traffic into statistically relevant information. Such information includes the most observed IP addresses from sending hosts, the most targeted ports on the receiving host, the type of internet traffic, times of traffic activity, geolocation of activity source, frequency of packets from any given IP address, and classifications of overall traffic activity [Harder et al., 2006]. Other relevant statistics include: number of unique source and destination ports; total number of IPv4 and IPv6 packets, as well as total number of bytes [Irwin, 2011]. These statistics will be represented in various formats, including geolocational graphs, time series plots, and tables of statistical analysis [Irwin, 2011].

4 Related Work

The analysis and reporting on of large datasets is by no means an original idea. Much mork has been done, both in the spheres of science and commerce, to mine extremely large datasets. One such example is the WEKA data mining software, which has been in development since 1993. The system aims to provide both machine learning algorithms and data preprocessing tools [Hall et al., 2009]. CAIDA, the cooperative association for internet data analysis, conduct regular measurements of Internet traffic across various networks [cai, 2014]. CAIDA also maintains the USCD network telescope, which observes traffic over a globally routed /8 network, as well as analysing the data collected from these observations [cai, 2012]. CAIDA has also put together an Educational Data Kit that uses data collected from the USCD telescope [Zseby et al., 2014]. PREDICT, the Protected Repository for the Defense of Infrastructure Against Cyber Threats, serves as a large repository of regularly updated network operations data-sets [PRE, 2014]. PREDICT is limited though as the restrictive access policies require researchers to work within certain geographical constraints, such as the continental United States [Irwin, 2011].

5 Research Approach

The first aim in the implementation is to develop a series of scripts to process the captures in the data-sets. The second aim is to build a reporting framework for the selection and formatting of the generated output based

on specialist scripts. From there the focus will be on making the framework more flexible and customisable with regards to the analysis and reporting of different data-sets.

6 Requirements and resources

One obvious requirement is access to the datasets created by network telescopes during their surveilance of unused IP space. Useful resources include python libraries that target data mining and analysis. Other useful resources, depending on the eventual direction of the project, could extend as far as space on a web server from which to host a web-based reporting engine. Server space will have to be allocated to host the project report website as well.

7 Progression Timeline

- Proposal Document - 3 March 2014
- Seminar on project - 4 March 2014
- Gather relevant articles, conference proceedings.
- Start implementation
- Literature Review - 30 May 2014
- Seminar - Early August 2014
- Website completion - 7 November 2014
- Project completion - Mid November 2014

References

The USCD network telescope. Online, August 2012. URL `http://www.caida.org/projects/network_telescope/`.

The PREDICT repository. Online, 2014. URL `https://www.predict.org/`. https://www.predict.org/.

Traffic analysis research. Online, February 2014. URL {`http://www.caida.org/research/traffic-analysis/`}.

H. Du and J. Yang. *Discovering Collaborative Cyber Attack Patterns Using Social Network Analysis*. Social Computing, Behavioral-Cultural Modeling and Prediction. Springer Berlin Heidelberg, March 2011.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutermann, and H. Witten, I. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, June 2009.

Uli Harder, Matt W. Johnson, Jeremy T. Bradley, and William J. Knottenbelt. Observing internet worm and virus attacks with a small network telescope. *Electronic Notes in Theoretical Computer Scinece*, 151(3):47 – 59, May 2006.

Barry Irwin. *A FRAMEWORK FOR THE APPLICATION OF NETWORK TELESCOPE SENSORS IN A GLOBAL IP NETWORK*. PhD thesis, Rhodes University, January 2011.

Barry Irwin. A baseline study of potentially malicious activity across five network telescopes. In *5th International Conference on Cyber Conflict*, 2013.

David Moore, Colleen Shannon, Geoffrey M Voelker, and Stefan Savage. Network Telescopes: Technical Report. Technical report, Cooperative Association for Internet Data Analysis (CAIDA), Jul 2004.

Ruoming Pang, Vinod Yegneswaran, Paul Barford, Vern Paxson, and Larry Peterson. Characteristics of internet background radiation. In *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, pages 27–40. ACM, 2004.

Tanja Zseby, Alistair King, Marina Fomenkov, and KC Claffy. Analysis of unidirectional ip traffic to darkspace with an educational data kit. 2014.