2019-12

Cyber System Assurance through Improved Network Anomaly Modeling and Detection

Bollmann, Chad A.

Monterey, California: Naval Postgraduate School

NPS NRP Executive Summary
Cyber Assurance through Improved Network Anomaly Modeling and Detection
Report Date: 10/04/19 Project Number (IREF ID): NPS-19-N039-A
Naval Postgraduate School Graduate School of Engineering and Applied Sciences

MONTEREY, CALIFORNIA

CYBER ASSURANCE THROUGH IMPROVED NETWORK ANOMALY MODELING AND DETECTION

Period of Performance: 10/01/2018–10/1/2019

Researchers:
Principal Investigator (PI): CDR Chad Bollmann, Ph.D., Graduate School of Engineering and Applied Sciences (GSEAS), Electrical & Computer Engineering (ECE)
Additional Researcher(s):
Mr. Jorge Gonzalez, Ph.D. candidate, Florida Atlantic University, Mathematics
Mr. Joshua Clymer, Naval Research Engineering Internship Program

Prepared for:
Topic Sponsor Lead Organization: N8
Topic Sponsor Name: N81IO Information Warfare Branch, LCDR Khoa Nguyen, USN
Topic Sponsor Contact Information: khoa.h.nguyen@navy.mil

NPS NRP Executive Summary

Cyber Assurance through Improved Network Anomaly Modeling and Detection

Report Date: 10/04/19 Project Number (IREF ID): NPS-19-N039-A

Naval Postgraduate School Graduate School of Engineering and Applied Sciences

EXECUTIVE SUMMARY

Project Summary

The objectives of this work were to investigate the source of the dual natures of network traffic (i.e., Gaussian and alpha-stable) in order prove the merit of further development, improvement, and application of non-parametric and parametric, alpha-stable network anomaly detectors. The results of this portion of the study have been accepted for publication (Gonzalez, Clymer, & Bollmann, 2019). The hypothesis behind the research is that alpha-stable models more accurately describe network traffic than Gaussian models, thus alpha-stable detection methods should provide superior accuracy over Gaussian approaches. Additionally, the increased precision of alpha-stable test statistics may permit differentiating between (i.e., classifying) different types of anomalies and cyber attacks. Finally, the increased precision of the studied traffic models and detectors should also permit more rapid detection of anomalies, enabling faster network defender response to adversary action.

The ultimate intention of this work is to refine the developed tools and test them on Department of Defense (DoD) networks, and enter into cooperative research and development agreements with cyber-security firms that defend DoD networks. This research makes progress towards addressing current network defense gaps identified by the research sponsor, particularly in the process step of Detect (part of the cyber response framework of Identify, Detect, Protect, React, and Restore).

Keywords: *alpha-stable, network anomaly detection, renewal theory, generalized central limit theorem, GCLT*

Background

It is accepted that some aspects of aggregated network traffic (e.g., inter-arrival times, packet and byte volumes per unit time) can be more accurately characterized using heavy-tailed models (Bollmann, Tummala, McEachen, Scrofani, & Kragh, 2018, p. 5524; Paxson & Floyd, 1995, p. 226; Simmross-Wattenberg et al., 2011, p. 494). This acceptance is complicated by observations that in smaller networks or under certain traffic conditions, the same features instead appear to possess exponential-tails (i.e., become Gaussian). However, little theoretical explanation exists for the co-existence of alpha-stable- and Gaussian-distributed traffic.

If aspects of network traffic are heavy-tailed, non-parametric (i.e., distribution-agnostic) test statistics as well as parametric test statistics (which require the presumption of a specific distribution) developed for heavy-tailed inputs, should demonstrate superior performance when compared to inappropriate parametric tests such as mean and variance. The principle investigator (PI) has previously shown that non-parametric, alpha-stable test statistics outperform Gaussian tests (Bollmann, 2018), but combinations of parametric test statistics have not been evaluated.

NPS NRP Executive Summary

Cyber Assurance through Improved Network Anomaly Modeling and Detection

Report Date: 10/04/19 Project Number (IREF ID): NPS-19-N039-A

Naval Postgraduate School Graduate School of Engineering and Applied Sciences

Findings and Conclusions

This study began with a literature review of network anomaly detection work grounded in heavy-tailed processes; no significant published works were found since the PI's dissertation literature review. An additional literature review into the history of alpha-stable network traffic theory was completed, and a series of works linking self-similarity, long-range dependence, heavy-tailed traffic characteristics, and alpha-stable distribution theory were identified.

From the assumption that network communications from an individual device can be treated as independent, identically-distributed (IID) random processes, two proven literature models support the aggregation of these IID processes to an alpha-stable result. The first, impulsive-based model, via the Generalized Central Limit Theorem (GCLT), demonstrates that IID processes of infinite variance will produce an alpha-stable result. When the individual IID processes exhibit finite variance, which can occur when devices and processes are relatively similar, a Gaussian aggregation results. The second aggregation model, based on renewal theory (Taqqu & Levy, 1986, p.73) that individual traffic impulses possessing varying "on" and "off" times can similarly aggregate to alpha-stable or Gaussian processes.

Two simple modeling algorithms were developed to validate the aggregation theories. First, individual device impulses (i.e., inputs) were cataloged from five seconds of four different network traffic traces based on source and destination internet protocol addresses and port numbers. These impulse catalogs were then used to reproduce (i.e., model) the source traffic and compared to the original traffic distributions using cumulative distribution function (CDF) plots and Kolmogorov-Smirnov (KS) test values as similarity measures. The KS test values indicate that the renewal model is slightly more accurate than the impulse model, but both models satisfactorily predict the original distribution. This conclusion is particularly promising because time constraints prevented significant refinement of the models.

The theoretical support of the impulse and renewal models for either Gaussian or alpha-stable distributions validates our hypothesis that alpha-stable models more accurately describe network traffic. Because the Gaussian is a special case of the alpha-stable distribution (where the tail parameter $\alpha = 2$), the alpha-stable distribution becomes the only logical choice for network traffic models and, by implication, anomaly detection test statistics.

Testing parametric, alpha-stable derived anomaly detection statistics was the focus of the second portion of our work. The parametric test results validate our hypothesis that alpha-stable test statistics outperform Gaussian tests, and that parametric tests slightly outperform non-parametric tests. One important finding is that the alpha-stable accuracy improvement margin is greater at lower false alarm rates; for any statistical implementation to be commercially viable, detection thresholds must be set such that false alarm rates less than one percent.

The period of performance for this project concluded before additional traces could be evaluated and all intended detector implementations could be evaluated. An eventual goal of this work is to use the labeled data and alpha-stable fits as input to a machine learning detection approach; preliminary qualitative

evaluation has shown that the results would be well-suited to a clustering approach due to easily-identifiable separation boundaries between benign and attack traffic cases.

As previously discussed, the overall results from this study integrate nicely with the existing body of work describing Internet traffic as self-similar and long-range dependent with scaling properties; these characteristics all relate to or result from alpha-stable network traffic. The characterization of traffic inputs as impulsive provides an intuitive method of using established renewal theory to show that aggregated traffic should approach Gaussian or alpha-stable limits. Additionally, this study is the first to propose an explanatory method of predicting aggregated traffic distributions and characteristics that have long been determined only empirically. The long-term implications of this study are that the accuracy of network traffic measurement and analysis may be significantly improved through utilization of existing alpha-stable-based approaches, as well as development and application of novel alpha-stable methods, particularly with respect to machine learning. Note: the data used in this work is evaluated as highly reliable and reflective of real-world network traffic.

Recommendations for Further Research

Several areas of this research topic warrant additional exploration. First, both the impulse and renewal process models should be assessed both qualitatively and quantitatively against network traffic traces consisting of different numbers and types of devices. The extensibility of this model to wireless network traffic also bears investigation, as different protocols facilitating the multiple access environment will alter the ON-OFF times of device transmissions. Examination of the quantitative results would permit selecting either the impulse or renewal model as the frequently "best" alternative for future research and development. Once validated, fingerprinting methods should be evaluated for the renewal process model with a goal of identifying optimum quantities of stored network trace data. Development of a fingerprinting method could permit saving high-quality, reduced-volume samples (i.e., snapshots) of network traffic sources and aggregated traffic that could serve to inform long-term models and anomaly detection systems.

Next, the ensemble anomaly detection systems, consisting of combinations of both parametric and non-parametric alpha-stable test statistics can be developed, evaluated, and optimized. These systems could integrate Gaussian test statistics in order to reduce computational overhead and detection delays when the traffic modeling and evaluation system senses Gaussian traffic characteristics. Finally, the ability of the detection system to rapidly detect anomalies should be rigorously examined; the resistance of the alpha-stable methodology to outlier skew implies that anomalies could be reliably identified using a minimum of subsequent samples. These minimum-repeatability tests would significantly improve overall detection system performance and system response times to denial-of-service attacks, improving network resiliency.

References

Bollmann, C. (2018). *Network Anomaly Detection with Stable Distributions*. Ph.D. dissertation submitted for publication.

Bollmann, C., Tummala, M., McEachen, J., Scrofani, J., & Kragh, M. (2018). Techniques to improve stable distribution modeling of network traffic. *Proceedings of the 51st Hawaii International Conference on System Sciences*, 5524–5531. http://hdl.handle.net/10125/50578

Gonzalez, J., Clymer, J., & Bollmann, C. (2019). Aggregated Impulses: Towards explanatory models for self-similar alpha stable network traffic. Paper presented at the 13th International Conference on Signal Processing and Communication Systems, Surfer's Paradise, Queensland, Australia.

Paxson, V., & Floyd, S. (1995). Wide area traffic: the failure of Poisson modeling. *IEEE/ACM Transactions on networking*, *3*(3), 226-244.

Simmross-Wattenberg, F., Asensio-Perez, J. I., Casaseca-de-la-Higuera, P., Martin-Fernandez, M., Dimitriadis, I. A., & Alberola-Lopez, C. (2011). Anomaly Detection in Network Traffic Based on Statistical Inference and α-Stable Modeling. *IEEE Transactions on Dependable and Secure Computing*, *8*(4), 494-509.

Taqqu, M. S., & Levy, J. B. (1986). Using renewal processes to generate long-range dependence and high variability. In *Dependence in probability and statistics* (pp. 73-89). Birkhäuser, Boston, MA.

Acronyms

Generalized Central Limit Theorem	GCLT
independent, identically-distributed	IID
Kolmogorov-Smirnov	KS
principle investigator	PI