

Towards Systematically Evaluating Flow-level Anomaly Detection Mechanisms

Daniela Brauckhoff
ETH Zurich, Switzerland
brauckhoff@tik.ee.ethz.ch

Ulrich Fiedler
ETH Zurich, Switzerland
fiedler@tik.ee.ethz.ch

Bernhard Plattner
ETH Zurich, Switzerland
plattner@tik.ee.ethz.ch

Abstract— Currently, flow-level anomaly detection systems get widely deployed in ISP networks to provide fast detection in case of large-scale anomalies such as worms, denial-of-service attacks, or flash crowds. Unfortunately, *benchmark evaluation traces* which would allow for systematically evaluating these anomaly detection systems are not available to neither research nor industry.

In this paper, we identify three major problems that hinder a systematic evaluation of flow-level anomaly detection systems. (1) Only very few backbone traffic traces are available to the research community due to privacy concerns of ISPs and their customers. (2) Available traces do not contain anomalies of varying intensities which are required for assessing the sensitivity of anomaly detection systems. And (3) available traces do not contain annotated anomalies, also referred to as ground truth. We discuss existing approaches that aim at overcoming these three problems, and identify their drawbacks.

We propose an alternative approach for generating benchmark evaluation traces, namely *synthetic generation of flow-level traffic traces*, and discuss why and how this approach can provide a solution to the identified problems. The two main challenges with such an approach are to define normal and anomalous network behavior, and to find realistic models describing normal and anomalous traffic at the flow level. We discuss our ideas for defining normal and anomalous traffic, and specify the framework for a novel flow traffic model targeted at anomaly detection. Finally, we provide an initial design for a synthetic flow trace generator.

I. INTRODUCTION

Recent events have shown that benign and malicious anomalies such as flash crowds, network outages, worms, and denial-of-service attacks have the potential to disrupt critical services and infrastructures. Motivated by the observation that detection at the network edge is not well-suited for containing such large-scale attacks, several anomaly detection systems for backbone networks have been developed. These systems operate on data aggregated at the flow level, e.g., using Cisco Netflow [15], since inspecting single packets is not feasible on high-speed backbone links.

Backbone anomaly detection systems come in two flavors: *anomaly-specific detection systems* such as [9], [16], [4], and more *holistic anomaly-diagnosing systems* such as [6]. In general, anomaly detection systems work by building a baseline model for normal network operation, based either on a training period or on expert-knowledge. The observed behavior of a system is then sequentially compared to this baseline model; and if the deviation between observation and baseline

in one interval exceeds a certain threshold, this interval is flagged as anomalous.

So far many different algorithms have been proposed for building the baseline model and for computing the deviation. The diversity of metrics used for anomaly detection, however, is rather limited due to the fact that flow records provide only a restricted set of information. Metrics which simply count observed traffic features such as bytes, packets, flows, unique IP addresses or ports are most common today; but recently, also more complex metrics based on distributions of traffic features such as entropy have been proposed.

In contrast to the vast amount of anomaly detection systems proposed, the effectiveness of these systems has not been well-investigated. Today, common practice for evaluating anomaly detection systems is to show that a system is capable of detecting a few specific anomalies using one or very few traffic traces. *Such simplistic evaluations, however, are not suitable to answer more general questions like:* Which metric is best for detecting a specific anomaly? Which components of normal traffic are responsible for false positives? How sensitive is a specific detection system, and what is the penalty of increasing its sensitivity? Are there general rules for identifying the optimum threshold for a specific metric or detection system?

We argue that in order to answer these questions, systematic evaluations based on benchmark traffic traces are required. Unfortunately, such benchmark traffic traces are not available today. We see mainly three unresolved research problems which are responsible for this situation:

- *Privacy-concerns:* Traffic traces are not made available to the research community due to privacy-concerns of ISPs and their customers. Existing approaches for anonymizing privacy-sensitive information in the traces, have not been successful in overcoming this problem.
- *Anomaly variability:* The few available traces contain only a limited set of anomalies with fixed intensities. Current approaches of anomaly injection in existing traces, or scaling of existing anomalies in traces provide only a partial solution to this problem.
- *Ground Truth:* Available traces do not contain annotated anomalies, also referred to as ground truth. Current approaches such as manual anomaly labeling or labeling through reference systems provide only limited or biased ground truth.

Since anonymization, modification, and annotation of real

traffic traces fail in generating benchmark evaluation traces, there is a clear need for an alternative approach. We believe that synthetic generation of benchmark traces has the potential to address the three identified problems. We envision a synthetic trace generation system that produces normal flow traffic according to a *baseline model*, and anomalous flow traffic based on a variety of *anomaly models*. However, the challenge with generating synthetic flow-level traces is twofold: Firstly, we need to define what is considered normal and anomalous network behavior, and secondly we need to find the appropriate normal and anomalous traffic models. We propose an event-driven approach for defining normal and anomalous network behavior, and specify the framework for a novel flow-based network traffic model targeted at anomaly detection.

The rest of the paper is structured as follows. In Section II we discuss drawbacks of current approaches that try to address privacy-concerns, anomaly variety, and ground truth in real traces. In Section III, we introduce our novel approach for generating benchmark evaluation traces. In Section IV, we discuss related work, and, finally, in Section V we will conclude.

II. DISCUSSION OF CURRENT APPROACHES

In this Section we introduce current approaches which aim at addressing privacy-issues, anomaly variability, and ground truth in real traces. We argue that the proposed approaches are not capable of completely solving the identified issues, and thus are not viable candidates for generating benchmark flow traces for evaluating backbone anomaly detection systems.

A. Approaches addressing privacy-concerns

The most obvious approach for studying anomalies would be to conduct measurements in real backbone networks since these provide diverse network environments as well as a diverse set of anomalies. The problem with traffic traces from commercial networks, however, is that they are not made available to the research community due to privacy concerns of ISPs and their customers. Consequently, only very few researchers have access to traces of commercial networks, and those researchers that have access are limited in the research they can conduct and the results they can publish. Besides that, there are a few research networks (most notably [1] and [2]) which make traffic traces available to the research community. However, these available traffic traces are not complete: they are sampled at very high rates (such as 1 out of 1000) and often additionally anonymized. Hence, they are only of limited use for systematically evaluating anomaly detection systems. Moreover, it is known that traffic characteristics in research networks do not necessarily reflect those of real networks.

Anonymization is the most common approach to deal with privacy concerns in any real-world dataset. In flow-level data, it is primarily the IP addresses which can reveal privacy-sensitive information. Different approaches for IP address anonymization in traffic logs, including novel methods which allow for prefix-preserving anonymization of IP addresses, are discussed in [10]. However, even when IP addresses

are perfectly anonymized, flow-level data still reveals other information - which can be used by adversaries for attacks, or which is of interest for competitors. Preventing this leak of sensitive information requires even further anonymization of traffic traces. The problem with further anonymizing data is, however, that this not only removes privacy- or business-sensitive information from the traces but also information which is important and valuable for research. Optimizing this anonymization trade-off is still an open research issue [11], [7]. Finally, the hardest challenge will be to convince trace owners of the necessity to share data, and that data sharing induces no business or legal risks to them.

B. Approaches providing anomaly variability

For evaluating the detection capabilities of a system over a wider range of network and anomaly scenarios, traces which contain anomalies of varying intensities are required. The intention behind is that, in the worst case, a system which is perfectly capable of detecting an anomaly of a certain intensity might fail completely at lower anomaly intensities. Additionally, for evaluating holistic anomaly detection systems targeted at diagnosing anomalies of different types, traces which contain a variety of anomalies are required.

Today, research focuses mainly on one approach to solve the problem of anomaly diversity, namely, modification of existing traces [8], [14], [6]. Anomalies are either injected in existing traces of normal traffic, or anomalies in existing traces are amplified or attenuated. However, the degree of diversity gained by this kind of modification is rather limited. Moreover, a simple additive injection, or multiplicative modification of anomalies completely disregards the impact of an anomaly on normal network traffic [8]. For example, extracting an existing worm scan anomaly from normal traffic and multiplying it by a distinct factor, does not account for the reduction in normal traffic that is induced by the scanning anomaly. Consequently, simple injection and modification approaches are not well suited for increasing the anomaly diversity in traffic traces.

C. Approaches for establishing ground truth

For systematically evaluating anomaly detection systems, perfect knowledge of all anomalies in the evaluation traces is required. More specifically, labeling of all traffic in a given trace as either normal or anomalous is a prerequisite for determining the false positive and false negative rates of anomaly detection systems. Today, there exist mainly two approaches for achieving ground truth: *manual anomaly labeling*, and *labeling through reference systems*.

Although widely applied due to a lack of better alternatives, manual labeling is not an appropriate method to achieve ground truth. There are mainly three problems with manual labeling of traffic traces: 1) it is biased by the knowledge and view of the person that is labeling the trace and thus not objective; 2) it is error-prone like every manual process; and 3) it is not repeatable unless the labeling process is documented in a very detailed manner.

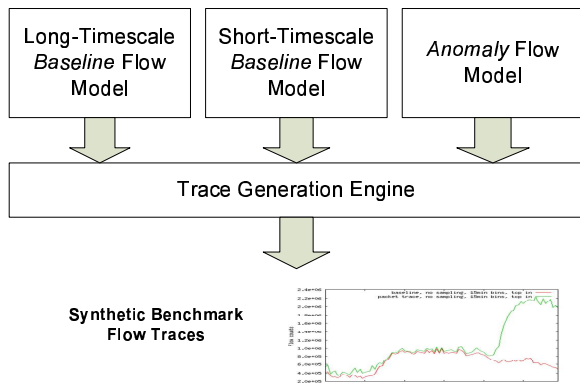


Fig. 1. Overview of the synthetic flow trace generation process. Flow traffic is generated by the trace generation engine according to a long/short-timescale baseline model, and in case of an anomaly, additionally according to an anomaly flow model.

The second common approach to achieve ground truth in traces is anomaly labeling through reference systems. The main problem with this approach is that it allows only for a comparative evaluation (the system under evaluation performs better or worse than the reference system) but not for an objective evaluation of a particular detection system. Moreover, there is no thoroughly evaluated anomaly detection system that could serve as a reference system today.

III. SYNTHETIC GENERATION OF BENCHMARK TRACES

In the last section we have provided arguments which underline that current approaches are not capable of making evaluation traces for flow-level anomaly detection systems widely available. Instead, we propose an alternative approach to address privacy-concerns, anomaly variability, and ground truth in evaluation traces: Synthetic generation of benchmark traces with the desired characteristics according to a flow-level traffic model. This section discusses the advantages and challenges of our approach.

In Figure 1, we present an overview of the synthetic flow generation process. In general terms, synthetic generation of benchmark traces is based on a *baseline model* which describes normal aggregated traffic at the flow level (at long-timescales as well as short-timescales), and different *anomaly models* which describe anomalous aggregate flow traffic. Flow traffic with annotated anomalies is generated by a *trace generation engine* according to these models.

A. Privacy and anomaly variability in synthetic traces

We argue that synthetic trace generation is a viable approach for generating benchmark traffic traces with two key characteristics: (1) They do not include any privacy- or business-sensitive information related to a particular ISP; and (2) they contain anomalies of different types and intensities.

Specifically, we believe that synthetic traces do not reveal any sensitive information about actual attacks since normal traffic, as well as anomalies, are generated purposely from an abstract traffic model. Moreover, the traffic model can be parameterized to reflect common network scenarios (e.g., a

middle-sized backbone network with several customer and peer PoPs), instead of reflecting the conditions of one specific ISP network. As a consequence, synthetic traces generated from such an abstract traffic model do not contain any privacy- or business-sensitive information and can thus be made available to the research community. Furthermore, anomalies of different types and intensities can easily be generated in a synthetic manner given that the underlying traffic models for normal and anomalous traffic provide the necessary parameters to allow such variations. To summarize, in order to generate synthetic benchmark traces with the characteristics specified above, the underlying traffic models must provide versatile means for parameterizing normal as well as anomalous flow traffic.

B. Ground truth in synthetic traces

While privacy and anomaly variability are addressed by selecting parameterizable traffic models for synthetic trace generation, the issue of ground truth is more difficult to solve. Since anomalies in synthetic traces are generated on purpose according to anomaly traffic models, ground truth should basically be known a priori. The problem, however, is that there is no widely agreed-on notion of normal and anomalous network behavior. But generating synthetic traces containing annotated anomalies implicitly requires a *reference system* which defines what is normal and what is anomalous network traffic.

There exist basically two approaches for defining anomalous network behavior. A *statistical approach* that treats outliers (in some measured metrics) as anomalies, and a more pragmatic *event-driven approach* which treats unusual events such as network outages, worms, or denial-of-service attacks as anomalies. However, both approaches do not qualify as reference system for the following reasons: The problem with the statistical approach is that it defines anomalies relative to normal behavior as captured by a specific metric. In contrast, the event-driven approach gives absolute definitions. However, there are other problems with the event-driven approach: it allows only for very general, descriptive anomaly definitions which cannot easily be transformed in a traffic model; it does not consider whether an event has an impact on normal network operation or not; and it defines a restricted set of anomalies.

Nevertheless, we prefer the pragmatic event-driven approach over the statistical one; mainly because the statistical outlier definition is not generic with regard to different detection metrics, temporal and spatial aggregation levels, as well as background traffic scenarios. To start with, we will define an explicit set of anomalies for which we provide a detailed definition and traffic model. We are aware of the restriction that this approach induces: We will only be capable of generating traces containing instances of this restricted set of anomalies. However, this set is not limited and can be extended over time. Moreover, normal network behavior is implicitly defined with this approach as being everything that is not labeled as anomalous.

C. Modeling Normal and Anomalous Traffic at the Flow Level

Having defined normal and anomalous network behavior, the next step is to find appropriate models that provide a realistic description of normal and anomalous traffic. A common guideline in traffic modeling states that each model should be designed for a specific purpose, in our case this is generating benchmark traces for flow-level anomaly detection. This specific purpose adds the following application-specific requirements to the general model requirements of parsimony, transparency, and generality:

- *Timescale of traffic model*: The timescale that anomaly detection systems operate at ranges from several days (detection systems are trained on several days of data) to minutes (detection metrics are aggregated over intervals of several minutes). Hence, our model must be capable of describing the long-timescale behavior (up to one week is presumably long enough), as well as the short-timescale behavior (down to one minute) of network traffic.
- *Flow characteristics to be modeled*: Generation of flow traces requires modeling of different flow parameters which are frequently used for anomaly detection: volume metrics use parameters such as packets and bytes, and spatial metrics use parameters such as source and destination addresses and ports.
- *Realistic and versatile anomaly models*: For generating benchmark traffic traces, the model must be capable of generating anomalies of varying intensities, and it must consider the impact of an anomaly on normal network traffic.

IV. RELATED WORK

Although several flow-based traffic models have been developed in previous work, we are not aware of a traffic model that meets all requirements specified in the last paragraph. The flow models proposed in [14], [3], [5], [12] all describe *volume* flow parameters, but they completely disregard *spatial* flow parameters such as IP addresses and ports. Consequently, these models are not suitable for evaluating anomaly detection systems which apply spatial aggregation metrics such as entropy. Additionally, several of the proposed models concentrate on describing the short-term behavior (at timescales of less than a minute) of flows. This contradicts our first requirement.

Furthermore, the flow model proposed in [14] is the only model that was designed for the specific purpose of testing anomaly detection systems. However, their proposed methodology for generating synthetic volume anomalies falls short in considering the interaction between anomalous and benign network traffic. Thus their model fails to meet the third requirement as specified above.

As already mentioned, similar problems exist with evaluating anomaly detection systems operating at the packet level. The authors of [13] applied synthetic generation of packet-level traces that contain certain anomalies for evaluating the performance intrusion detection systems. Their model, however, is targeted at generating packet-level characteristics of attacks instead of anomalous backbone traffic.

V. CONCLUSION

In this paper, we have identified three main challenges that hinder systematic evaluations of flow-level anomaly detection systems, namely privacy-concerns, anomaly variability, and ground truth in traffic traces. We address several drawbacks of current approaches that try to meet these challenges, and argue that these approaches are not suitable for generating benchmark evaluation traces.

Instead, we propose synthetic trace generation as an approach which has the potential of making benchmark traffic traces available to a broad community. We present a reference system that allows us to define normal and anomalous network behavior. Moreover, we define the requirements for a traffic model that is capable of generating synthetic benchmark evaluation traces with the desired properties. We discuss previously developed flow models, and argue that none of these models meets all specified requirements. Finally, we believe this work opens up new directions for systematically evaluating anomaly detection systems.

REFERENCES

- [1] Abilene network operations center. <http://www.abilene.iu.edu/>.
- [2] Géant2 - The pan-European research network. <http://www.geant2.net/>.
- [3] C. Barakat, P. Thiran, G. Iannaccone, C. Diot, and P. Owezarski. Modeling Internet backbone traffic at the flow level. *IEEE Transactions on Signal Processing*, 51, August 2003.
- [4] Thomas Dübendorfer and Bernhard Plattner. Host Behaviour Based Early Detection of Worm Outbreaks in Internet Backbones. In *WET ICE 2005, Linköping, Sweden*, 2005.
- [5] S. Ben Fredj, T. Bonald, A. Proutiere, G. Régnié, and J. W. Roberts. Statistical bandwidth sharing: A study of congestion at flow level. In *SIGCOMM '01: Proceedings of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications*, pages 111–122, New York, NY, USA, 2001. ACM Press.
- [6] Anukool Lakhina, Mark Crovella, and Christophe Diot. Mining Anomalies Using Traffic Feature Distributions. In *Proceedings of ACM SIGCOMM 2005*, August 2005.
- [7] Ruoming Pang, Mark Allman, Vern Paxson, and Jason Lee. The Devil and Packet Trace Anonymization. *ACM Computer Communication Review*, January 2006.
- [8] Andy Rupp, Holger Dreger, Anja Feldmann, and Robin Sommer. Packet trace manipulation framework for test labs. In *IMC '04: Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, pages 251–256, New York, NY, USA, 2004. ACM Press.
- [9] Vyas Sekar, Nick Duffield, Oliver Spatscheck, Jacobus van der Merwe, and Hui Zhang. LADS: Large-scale Automated DDoS Detection System. In *USENIX Annual Technical Conference*, 2006.
- [10] A. Slagell, Y. Li, and K. Luo. Sharing network logs for computer forensics: A new tool for the anonymization of netflow records. In *CNFR Workshop, held in conjunction with IEEE SecureCom, Athens, Greece*, 2005.
- [11] A. Slagell and W. Yurcik. Sharing Computer Network Logs for Security and Privacy: A Motivation for New Methodologies of Anonymization. In *SECOVAL, held in conjunction with IEEE SecureCom, Athens, Greece*, 2005.
- [12] Joel Sommers and Paul Barford. Self-configuring network traffic generation. In *Internet Measurement Conference*, pages 68–81, 2004.
- [13] Joel Sommers, Vinod Yegneswaran, and Paul Barford. A framework for malicious workload generation. In *Internet Measurement Conference*, pages 82–87, 2004.
- [14] Augustin Soule, Kavé Salamatian, and Nina Taft. Combining filtering and statistical methods for anomaly detection. In *IMC '05: Proceedings of the 5th ACM SIGCOMM conference on Internet measurement*, 2005.
- [15] Cisco Systems Inc. Netflow services and applications - white paper.
- [16] Arno Wagner and Bernhard Plattner. Entropy based worm and anomaly detection in fast ip networks. In *WET ICE 2005, Linköping, Sweden*, 2005.