Network Traffic Emulation for IDS Evaluation

Yang Wang, Gong Jian, Ding Wei and Wu Xiong
School of Computer Science and Engineering,
Southeast University, Nanjing, Jiangsu, China 210096
{wyang, jgong, wding, xwu}@njnet.edu.cn

Abstract

The Network traffic Emulation is used in generating background traffic for IDSs evaluation. The Background traffic can be used to evaluate the false positive level and the performance of the misuse IDSs and help training normal behavior profiles for anomaly IDSs. Currently the emulation methods for the background traffic are either restricted by the performance bottleneck of the software and hardware, or lack of the semantic of flow and session. So they can't satisfy the IDS evaluation requirement in high-speed network environment. After analyzing the requirement of IDSs evaluation and the characteristics of network traffic, this paper proposes a differential equation model of active flow rate. Based on the equation, a structural simulation model of network flow is constructed and used in the network traffic emulation for IDS evaluation. This model is both simple for high performance and similar to the reality. The experiments show that the model proposed can generate traffic both realistic and controllable

1. Introduction

The IDSs(Intrusion Detection Systems) are now standard equipments for large networks, and IDS evaluation is needed by the researcher. The background traffic is a important part in the IDS evaluation. Comparing to using real environment or the trace as the background, traffic emulation method have the characteristics of repeatable and informative. The repeatable allows for independent validation, the informative provides an understanding of the causes underlying performance behaviors, thereby facilitating improvements

In the follow section, we will describe related work and analyzes the requirement for network traffic emulation in high-speed network environment IDS evaluation. Section 3 introduces the active flow rate by differential equation model. In Section 4, the structural emulation model based on the differential equation is proposed. The experiments of performance and reality are analyzed in Section 5. Finally, concluding remarks are given and possible future work is discussed in Section 6.

2. Requirement Analysis

2.1. Related Research

MIT/LL has performed the most extensive quantitative IDS testing to date. Sponsored by the Department of Defense Advanced Research Projects Agency (DARPA), MIT/LL conducted large-scale testing of research IDSs in 1998 and 1999. Recent extensions of the MIT/LL evaluations led to an evaluation test-bed entitled LARIAT (Lincoln Adaptable Real-Time Information Assurance Test-bed) that automates most of the tasks required for real-time evaluations. MIT/LL used automata models to simulate different user behaviors. The models are implemented in scripts that droved applications interacting with servers to generate network traffic. This background generating method can provide the packets with semantic of transport and application level. But because of using real PC representing network nodes, this method can't simulate the topology and user's behavior of backbone network. And the method driving application is limited by the resource of the operating system. In 98/99 testing the traffic volume is less than 100kbps [6]. The LARIAT system traffic volume is under 100Mbps. Although MIT/LL's method is good for IDS test in small or medium organization network, this method can't fit the high volume traffic environment request.

Vendor independent Test labs like NSS Group use commercial broadband test tool to generate background traffic, such as Smartbits, Web Avalanche and Web Reflector [5]. Tools like Smartbits were

designed for load-testing switches and routers; they can't generate traffic with transport and application level semantic. So these systems can't reach the request of IDS evaluation.

In 2004 Joel Sommers[8] published a new application-independent tool Harpoon for generating representative packet traffic at the IP flow level. Harpoon use Net-flow records of routers to generate TCP and UDP packet flows. These flows have the same byte, packet, temporal and spatial characteristics as measured at routers in live environments. Its generating speed is about 600Mbps. But as a router benchmarking tool, Harpoon also lack of the application level semantic.

2.2. Traffic Emulation for IDS Evaluation

Traffic emulation can have different levels. Most IDSs now detect anomaly behavior based on sessions. Traditional packet level emulation methods can't have the semantic of the session, so they can't fulfill the requirement of evaluation. As the basic of sessions, flow emulation methods are needed. For different evaluation objects, flow emulation has different methods. Alexsson classified the IDSs as signature-based and anomaly-based. The signature-based IDSs' analytic object is payload (single packet payload, multi packet assembled payload), so traffic behaviors only have impaction on IDSs' performance. And fixed-rate traffic can be used in the stress test of these IDSs. For anomaly-based IDSs, traffic behaviors are also used by IDSs to learn and construct normal behavior profiles. So traffic behaviors are valuable for not only the stress test, but also the function test for the anomaly-based IDSs. Besides high-speed, the traffic emulation objects include ability to emulate traffic behaviors in different network environment with good scalability.

The Flow emulation includes different characteristics such as flow behaviors, topology and payload. This paper concerned on the primary characteristic – flow behavior characteristic. It is the base of flow existence, and it is the base of emulation model.

Flow behavior characteristic is the statistical laws of flow in time and spatial metrics. Time metrics include flow arrival rate, flow duration, flow shape (packet arrival rate in flow) and active flow rate etc. Flow arrival rate, flow shape is directly emulated metrics, and flow duration, active flow rate are indirectly emulated metrics. Spatial metrics include flow length (packet number in a flow) and flow throughput (byte number in a flow). Flow length is directly emulated metric, and flow throughput is indirectly emulated metric. Paper [1] present that active flow may be considered the real network state and is presumably a better indication of utilization or desired operating point. So in third section the active flow rate metric is selected as emulation object. Through analysis of flow behavior metrics, active flow rate metric is represented as a differential equation of directly emulated metrics.

3. Flow Structural Model

The NIDS cares about the network traffic in its monitoring area. It doesn't care about how the packet travels from the source node to destination node. So the model object is the traffic in the monitored network. In the modeling process, the protocols behavior difference was ignored, and different protocol's flows are treated as abstract packet stream. In the implementation, different parameter was assigned to different protocols in order to distinguish their difference. So in the following chapter the behavior difference of different protocol was ignored, and all the protocols flows were seen as one kind of abstract flow.

Define M_{flow} as the mean value of flow arrival rate. In this model M_{flow} is used to count flow quantity. There are two assumptions when M_{flow} is used: 1) the first is flow arrival is a Poisson process; 2) the second is flow arrival rate is fast enough. Paper [1, 2] both proposed that simulating flow arrival can use Poisson process. If flow quantity is too little in time Δt, the mean value can't be use to count flow quantity.

Define $N_{\text{flow}}(t)$ as the number of active flow at time t. $N_{\text{flow}}(t)$ is the sum of the active flow number at last time ($t - \Delta t$) plus new flow number in Δt and sub finished flow number in Δt. Define $A_{\text{flow}}(t)$ as flow arrival rate, and define $L_{\text{flow}}(t)$ as flow leave rate, $N_{\text{flow}}(t)$ is:

$$N_{flow}(t+\Delta t)=N_{flow}(t)+A_{flow}(t)\cdot\Delta t-L_{flow}(t)\cdot\Delta t \ . \quad (1)$$

Transform the Equation 1 into the differential form:

$$\frac{dN_{flow}}{dt} = A_{flow}(t) - L_{flow}(t) \ . \quad (2)$$

Flow arrival rate can be represented by the mean value M_{flow}:

$$A_{flow}(t) = M_{flow} \ . \quad (3)$$

Flow leaving rate can be described by the flow duration and the flow arrival rate. Define $F_D(t)$ as

cumulative density function of flow duration, then flow leaving rate is:

$$L_{flow}(t) = \int_0^t M_{flow} \cdot f(x)dx = M_{flow} \cdot F_D(t) \ . \quad (4)$$

Now integral Equation 2 to Equation 5, and define $CF_D(t)$ is complementary cumulative density function (CCDF) of flow duration：

$$N_{flow}(T) = M_{flow} \cdot \int_0^T CF_D(t)dt \quad (5)$$

Equation 5 needs to be represented by the packet metric so as to emulation traffic can be generated. Because flow duration equals to the sum of packets arrival gap in a flow, we can get Equation 6:

$$D = \sum_{i=1}^{n_{flowlen}} g_i \quad (6)$$

Flow duration has two factors: flow length (the packet number of a flow $n_{flowlen}$) and packet arrival gap in a flow (g_i). Flow length is related to application, while packet arrival gap is related to operation network state, which means that they are relative independent. Considering operation network state, flow arrival gap can be expressed by formula $g_i = c + (L_f + L_r)/b + Q(t)$, c is transmitting delay, L_f is forward flow throughput, L_r is backward flow throughput, b is available bandwidth, and Q(t) is queuing delay of routers[2]. The latter two factors are small enough to be ignored. For the first factor transmitting delay, Baek-Young Choi [2] discovered that: 1) delay distributions vary greatly in shape, depending on the path and link utilization; 2) after constant factors dependent only on the path and packet size are removed, the 99th percentile variable delay remains under 1 ms over several hops and under link utilization below 90% on a bottleneck; 3) a very small number of packets experience very large delay in short bursts. So mean value or quantile can be used as packet arrival gap estimation. g_i in Equation 6 can be represented mean value or quantile G:

$$D = n_{flowlen} \times G \quad (7)$$

So complementary cumulative density function of flow duration $CF_D(t)$ can be replaced by complementary cumulative density function of flow length $CFn_{flowlen}(t)$:

$$N_{flow}(T) = M_{flow} \cdot \int_0^{\frac{T}{G}} CF_{n_{flowlen}}(l)dl \quad (8)$$

Equation 8 imported total three metrics: flow arrival rate, flow length and packet arrival gap. The three metrics are related to three levels: flow arrival rate is the characteristics of application level; flow length reflects the distribution of transmitted file, it is the characteristics of flow content; packet arrival gap is the result of interacting between transmit and network level, it's the transmission characteristics. According to the Monte Carlo method, the three metrics can be used to emulate the traffic with similar characteristics of real traffic. Section 5 compared CDF graph of emulated traffic and real traffic to prove the emulation method's effectiveness.

With the growth of Internet, there are a greater number of new application emergence, and old application's parameters are changing. [4] But the emulation model can adapt to the changing by substituting old distribution functions and parameters with new ones.

4. Network Traffic Emulation Model for the IDS Evaluation

Based on the active flow model in section 3, a three level structural emulation model (SEM) can be constructed. SEM has three levels: application /content /transmission. Each level has a variable: flow arrival rate S in application level, flow length L in content level, and packet arrival gap G in transmission level.

Section 3 use the mean value of flow arrival rate to count the active flow under the assumption flow arrival rate is a Poisson process. In order to be more similar to real traffic in the behavior, evaluation process uses Poisson process to emulate flow arrival process. In the future research, how to emulate the non-homogenous Poisson process will be studied.

The emulation algorithm based on SEM is:

1) Calculate flow arrival rate S and packet arrival gap G according to the test requirement.
2) Select flow length distribution L according to the test requirement.
3) Set emulation time T according to the test requirement, flow length distribution.
4) Use random number generator confirming with Poisson process to generate the arrival time of each flow according to flow arrival rate.
5) Use random number generator confirming with flow length distribution to generate length of each flow according to flow length distribution.
6) Generate packet arrival time for each packet

according to flow arrival time and packet arrival gap.

7) Fill each packet's payload according to the protocol content template.
8) Merge packets to the trace file or send directly through network interface.

5. Experiment and Verification

5.1. Performance Experiment

A prototype system was built based the model of section 4 for experiment. The system is coded in the C language. The random number generator is based on GSL lib (GNU Scientific Library) version 1.4. The Packet constructor is based on libnet library version 1.1. Figure 1 is the performance test result of the system.

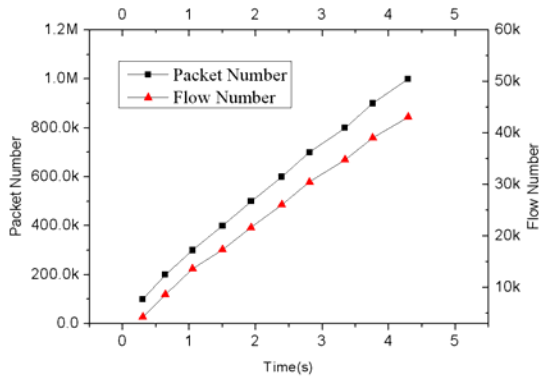

Figure 1. Traffic generate rate

The environment of experiment is a server with two XEON 2.4G CPU and 1G RAM. Average packet length is set as 470 bytes. There are ten groups of experiments. The packet number of experiments started from 100K and increased 100K packets each experiment. Figure 1 showed that the system generating speed is about 200kpps / 800Mbps and is linear. At the speed the system utilized 80 percent of 1Gbps link. So the system can effectively fulfill the performance request of 1Gbps link.

5.2. Reality Experiment

Abilene trace was used to verify the similar between emulation traffic and real traffic. Abilene-I, an OC48c Packet-over-SONET data set published by the NLANR (**N**ational **L**aboratory for **A**pplied **N**etwork **R**esearch) MNA team. Because there are anomaly traffic in the traces like attacks and route loop, they may disturb the characteristics of flow arrival rate. To guarantee the emulated traffic has the characteristics of Poisson process, the trace was filtered. The filter rules will described with the experiment.

In the experiment, 600 seconds traffic from the beginning of the trace is used as sample. The emulation process used the parameters derived from the 600 seconds sample. Then the emulation process generated 3000 seconds traffic and compare it with the real traffic.

Abilene-I consists of a pair of two hour contiguous bidirectional packet header traces collected at the Indianapolis router node (IPLS). The filter rules were: 1) flow length is at least ten packets; 2) the flows have complete interaction process, which mean flow started with SYN packet and ended with FIN or RST packet. First the flow length distribution function of trace need be fitted. The Pareto distribution was used as fitted function. Because After filtering flow length is at least 10 packets, position parameter of Pareto distribution is set as 10. When shape parameter of Pareto distribution is set as 0.9 and 1.0, the emulated length distribution is closest to the real flow length distribution. So in emulation process the flow length distribution used Pareto distribution, the shape parameter is set 0.9, 1.0 and 0.95. From the 600 seconds trace flow arrival rate mean value is 208.33flow/second and packet arrival gap is 0.23 second. Figure 2 showed the active flow rate of emulation traffic and real traffic. Figure 3 showed the active flow rate CDF of emulation traffic and real traffic.

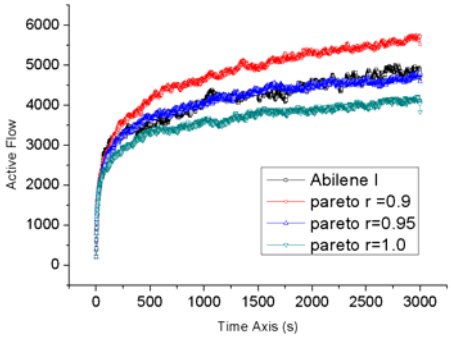

Figure 2. Active flow rate

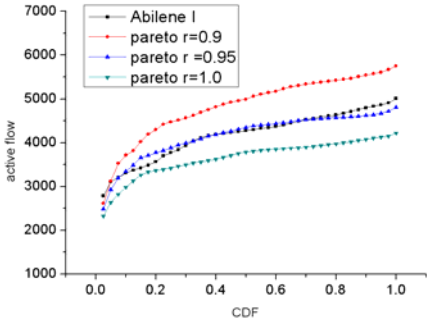

Figure 3. Active flow rate cdf

Figure 2 showed the active flow rate comparison between emulation and real traffic. Figure 3 showed the active flow rate CDF comparison between emulation and real traffic. In both figures, when r equals to 0.95, the emulation traffic is closest to the real traffic. When r equals to 0.9 and 10, the emulation traffics active flow rate curve s are equally on the both sides of the real traffic curve. For these trace, the emulation method effectively emulated the real traffic.

6. Conclusion and Future Work

The paper proposed a network traffic emulation method based on the structural model of network flow. The method can not only satisfy the requirement of high-speed network environment test, but also emulate the traffic behavior with high similarity. At First the active flow rate model is discussed. Then a structural emulation model is proposed on the active flow rate equation. The experiments proved the emulation method is better than traditional methods in performance and similarity to real traffic. This method can not only emulate normal high-volume traffic, but also be extended to emulate anomaly traffic like worm and route loop.

The model assumed that the IDSs monitor all the protocols with the same way. But in reality IDS may use different ways to monitor different protocols. The future work will care about more application protocols behavior, and extend the single dimension three-level structural model to two dimensions three-level structural model (application specified SEM). The new model will develop different models for different application level. Similarly, the topology and payload metrics will be considered of their impaction on the IDS evaluation as well as the behavior discussed in this paper.

7. Acknowledgment

This research is partially support by the National Basic Research Program (called 973 Program), No. 2003CB314803; Jiangsu Province Key Laboratory of Network and Information Security BM2003201 and the Key Project of Chinese Ministry of Education under Grant No.105084.

8. References

[1]. Afanasiev, F., Petrov, A., V.Grachev and Sukhov, "A. Flow-based analysis of Internet traffic". *Russian Edition of Network Computing*, *5* (98). 92-95.

2. Barakat, C., Thiran, P., Iannaccone, G., Diot, C. and Owezarski, P. Modeling Internet Backbone Traffic at the Flow Level. *IEEE TRANSACTIONS ON SIGNAL PROCESSING*, *51* (8). 2111-2224.

3. Choi, B.-Y., Moon, S., Zhang, Z.-L., Papagiannaki, K. and Diot, C., Analysis of Point-To-Point Packet Delay In an Operational Network. in *Infocom 2004*, (Hong Kong, 2004).

4. Floyd, S. and Paxson, V. Difficulties in Simulating the Internet. *IEEE/ACM* Transactions *on Networking*, *9* (4). 392 - 403.

5. *NSS* Group, Intrusion Detection Systems Group Test (Edition 4). NSS Group, 2004.

6. Haines, J., Lippmann, R., Fried, D., Korba, J. and Das, K. Design and Procedures of the 1999 DARPA Intrusion Detection Evaluation: Design and Procedures, MIT Lincoln Laboratory, 2001.

7. Hong, S.-S. and Wu, S.F., On Interactive Internet Traffic Replay. in *International Symposium on Recent Advances in Intrusion Detection (RAID 2005)*, (Seattle, Washington, USA, 2005), 247-264.

8. Joel, S. and Paul, B. Self-configuring network traffic generation *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, ACM Press, Taormina, Sicily, Italy, 2004, 68-81.

9. Mchugh, J. Testing Intrusion Detection Systems: A Critique of the 1998 and 1999 DARPA Intrusion Detection System Evaluations as Performed by Lincoln Laboratory. *ACM Transactions on Information and System Security*, *3* (4). 262-294.