Anomaly Detection Method based on Pattern Recognition

Romain Fontugne
The Graduate University for Advanced Studies
romain@nii.ac.jp

Yosuke Himura
The University of Tokyo
him@hongo.wide.ad.jp

Kensuke Fukuda
National Institute of Informatics / PRESTO, JST
kensuke@nii.ac.jp

1. INTRODUCTION

Identifying anomalies in network traffic is an important task for securing operational networks and maintaining optimal network resources available. Recently, researchers have mainly tried to handle anomaly detection as a statistical problem, proposing several statistics-based methods [1, 4]. Because it is problematic to estimate significant statistics from small (mice) flows statistics-based methods have a common difficulty to analyze mice flows. We focus on detecting low-intensity traffics since sophisticated or large scale attacks tend to be distributed processes involving numerous hosts generating only mices.

We introduced a new approach based on pattern recognition [3] using network-related information. Anomalous traffics are emphasized through their excessive utilization of a few traffic features represented as "lines" in pictures monitoring the traffic. Anomalies are easily extracted with a line detector and original data are retrieved from identified plots. Main advantages of this method are its ability to report quickly and precisely anomalies involving a tiny amount of packets. The proposed method analyzes only header information of the traffic, and requires no prior information on the traffic or port numbers. Like some of recent anomaly detection methods [1, 4], the pattern-recognition-based method is able to identify long/short-lasting anomalies appearing simultaneously, and anomalies already started or not yet finished. However, it is distinguished from the other approaches by detecting anomalies represented by a tiny number of packets and by identifying precisely packets involved in anomalies.

2. TEMPORAL-SPATIAL BEHAVIOR

Here we focus on the manner to highlight anomalies through their unusual uses of network traffic features during a period of time. We consider a few traffic features (namely source address, destination address, port source, port destination) and demonstrate that anomalous traffics manifest abnormal distribution of some of them. By mapping traffic into a 2-D space (one feature and the time), anomalies are intuitively identifiable as "lines". Figure 1 illustrates several examples of anomalies identified in the same day (2004/10/14), legitimate traffic have been exclude for clarity. The blue lines in the left side of Fig. 1 are generated by one host probing a large sub-network on port 5900 (VNC). Green slanted lines stand for a similar behavior against a Windows service (port 445). Whereas the red horizontal lines display abnormally high traffic between a couple of hosts on port 8000. The analyzed traffic has been flanked by two significant outbursts of the Sasser worm. Sasser activity is monitored in black in Fig. 1, where two different propagations of the worm are shown. On the one hand long vertical lines are drawn on the whole picture depicting a quick and large spreading. On the other hand small slanted lines are shown on the top side of the figure standing for a slow spreading of the worm.

PAM2009, April 1-3, 2009, Seoul, Korea.

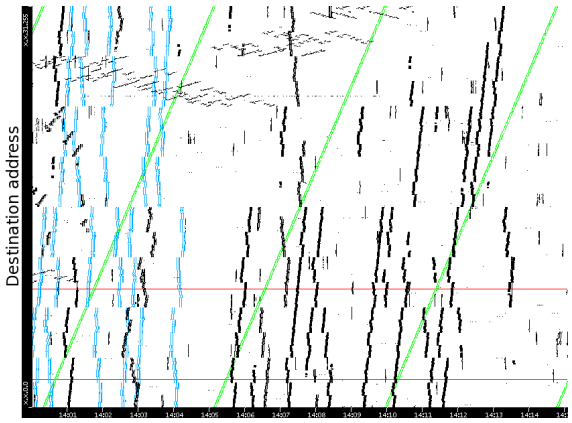

Figure 1: Several examples of anomalies identified in one traffic trace of 15 minutes (2004/10/14).

3. DETECTION METHOD

The core of the detection process consists of the three following steps:

- Computation of pictures: Four categories of picture are considered to emphasize anomalous traffic, all of them have the time on the x axis and a different traffic feature on the y axis (source/destination address or port). For each category several pictures are computed representing only sub-traffic aggregated jointly in function of their sources or destinations. Consequently, traffic behavior is represented in various pictures in which the noise is reduced and anomalies are highlighted.

- Detection: Hough transform: Our method is based on the Hough transform [2] to detect lines in pictures. We point out two important assets of this well-known technique used in pattern recognition: (1) It allows to de-

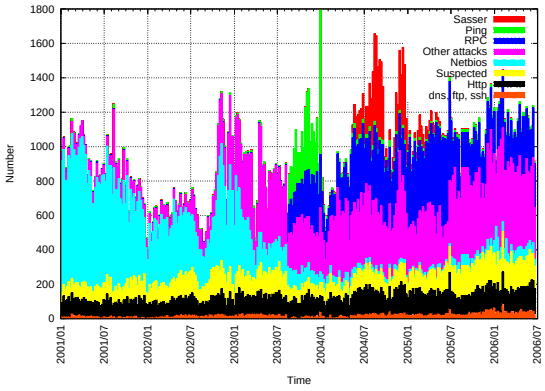

Figure 2: Anomalies reported by the proposed method on traffic traces collected in MAWI from 2001 to 2006.

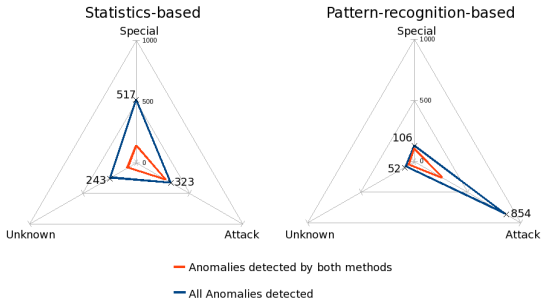

Figure 3: Anomalies identified by both methods (MAWI 2004/08/01).

tect lines with missing parts (e.g. dotted lines). Consequently anomalies displayed as segmented lines (due to network or process latency) are also identifiable. (2) It is robust against noise, thus it detects anomalies surrounded by dense legitimate traffics which generate noise on analyzed pictures. The Hough transform consists of a voting procedure, where each plotted point (x, y) of a picture elects lines that can pass through its position (x, y). Using the polar coordinates it consists in enumerating all ρ and θ solving the following equation: $\rho = x \cdot \cos\theta + y \cdot \sin\theta$. All votes are collected in an array called Hough space (or accumulator), and all candidate lines are determined as the maximum values in this accumulator.

- Identification: For each line extracted by the Hough transform initial data are recovered from the position of all plots involved. Packet information are summarized in a set of statistics called *event*. Thus, an *event* constitutes a report for a specific line emphasized in a picture. Anomalies are monitored by more than one line and raises several events. Therefore, *events* from the same source or aimed at the same destination are grouped together to form an *anomaly*.

4. EVALUATION

We tested our method on several traffic traces from the MAWI archive (see Fig. 2), and noticed that the proposed method is able to efficiently detect numerous types of anomalies. We compared our results with those given by a statistics-based method [1]. We deduced from this analysis that even though the two methods have different theoretical background, they had almost 50% of their results in common. Figure 3 shows the results of both methods splitted into three categories: Attack (well-known attacks), Special (common services: http, ftp, dns) and Unknown (mostly peer-to-peer).

The statistics-based method analyzed in depth characteristics of flows and detected effectively traffic with singularities, but the maliciousness of these traffics is often difficult to determine without packet payload, and analysis is not processed on small flows (to avoid a high rate of false positive alarms). The method based on pattern recognition identified clear anomalous behavior, and successfully detected two times more traffics related to worms and scans activities than the other method. This category of anomaly stands for small flows and represents the fundamental weakness of statistics-based methods. Consequently, the comparison reavealed that the two approaches detect classes of anomalies with different characteristics. These two methods have distinct weaknesses and advantages, thereby they are a good combination for anomaly detection.

5. CONCLUSION

We illustrated characteristic shapes of anomalous traffic in time and space, and presented an original approach for anomaly detection based on pattern recognition. This method takes advantage of graphical representation to reduce the dimensions of network traffic, and benefits of techniques form image analysis. Only header information is required, no inspection in the packet payload is processed and no prior information on the traffic or ports number is needed. Analyzing traffic from a trans-Pacific link reavealed that the method we proposed is able to identify various classes of anomalies, with a sensibility to worms and network/port scans, and its ability to detect tiny flow is a novelty in anomaly detection. The comparison of the proposed method and a statistics-based method indicates that the two approaches successfully identified distinct classes of anomalies. Therefore, the use of the alternative method proposed and a statistics-based method is a good combination and can provides synergy. One important future work is to compare our method with other detection methods on large dataset. Also, understanding the parameter set of the proposed method would allow automatical tuning of the tool.

6. REFERENCES

[1] G. Dewaele, K. Fukuda, P. Borgnat, P. Abry, and K. Cho. Extracting hidden anomalies using sketch and non gaussian multiresolution statistical detection procedures. *LSAD '07*, pages 145–152, 2007.

[2] R. O. Duda and P. E. Hart. Use of the hough transformation to detect lines and curves in pictures. *Commun. ACM*, 15(1):11–15, 1972.

[3] R. Fontugne, T. Hirotsu, and K. Fukuda. An image processing approach to traffic anomaly detection. *AINTEC '08*, pages 17–26, 2008.

[4] A. Lakhina, M. Crovella, and C. Diot. Mining anomalies using traffic feature distributions. *SIGCOMM '05*, pages 217–228, 2005.