Detecting Anomalies in Massive Traffic with Sketches

Sirikarn Pukkawanna, Hiroaki Hazeyama, Youki Kadobayashi, and Suguru Yamaguchi
Nara Institute of Science and Technology, Japan
{sirikarn-p, hiroa-ha, youki-k, suguru}@is.naist.jp

Motivation and Challenge

- Detecting network anomalies is crucial
  - Attacks, spreading of worms, outages
- Limitations of signature-based anomaly detectors
  - Need up-to-date attack signatures
  - Cannot detect unknown and new attacks

➢ Internet traffic data is exponentially growing and new attacks are constantly invented
➢ *Traffic analyzers that do not require prior knowledge as well as can handle the higher data rate are needed*

Proposed Three Steps to detect Anomalies in Massive Traffic

❶ Summarize traffic stream using sketches:

➢ Split traffic into several sub-stream by hash functions as shown in **Fig. 1**

➢ Compute Entropy of each sub-stream. Entropy is defined as $H(X) = -\sum_{i=0}^{n} p_i log_2 p_i$, where

$$p_i = \frac{\#pktsAssocWith\_srcIP_i|dstIP_i|srcPort_i|dstPort_i)}{\#totalPktsSeen}$$

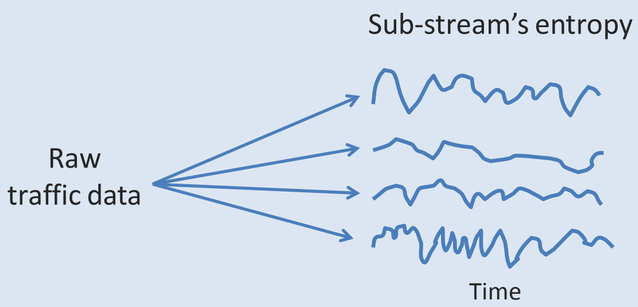

Fig. 1 Traffic summarization

❷ Detect time-bins that contain changes based on S-transform:

➢ S-transform converts the entropy to time-frequency domain as shown in **Fig. 2**

➢ Find changes in the time-frequency domain

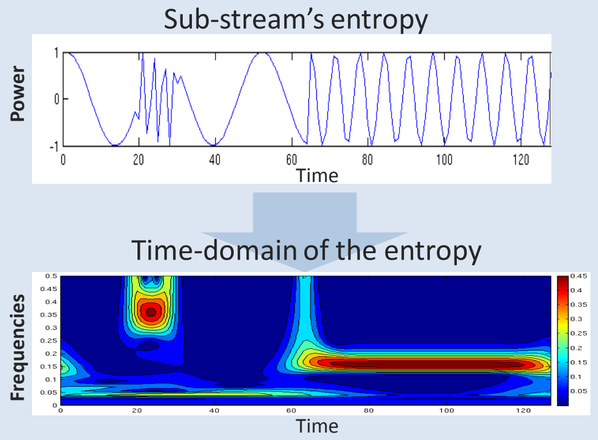

Fig. 2 Frequency extraction by S-transform

❸ Detect anomaly culprits:

➢ Find the keys (e.g., source IP) in the detected suspect time-bins

Evaluations with Real-world Backbone Traffic Collected at the 150 Mbps US-JP Link

Accuracy rate[1]

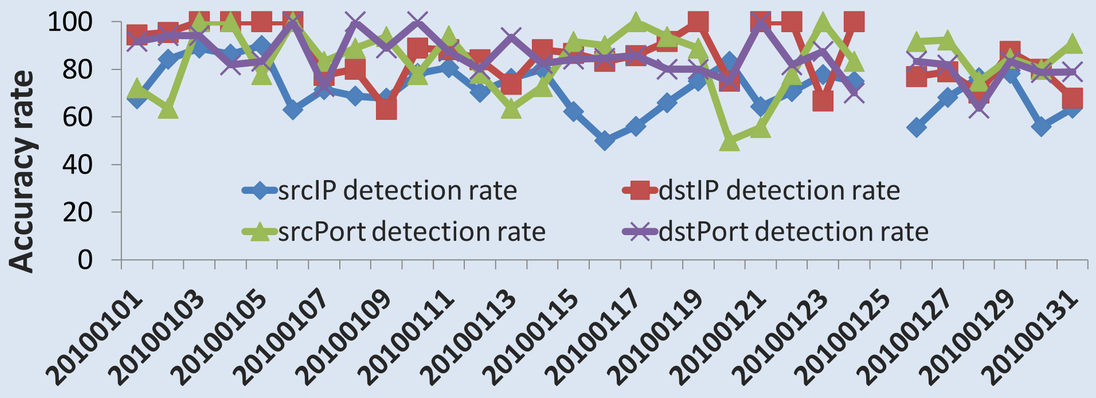

Fig. 3 Accuracy rate of detecting anomalous source IP, destination IP, source port, and destination port in traces collected on January 2010

False positive rate[2]

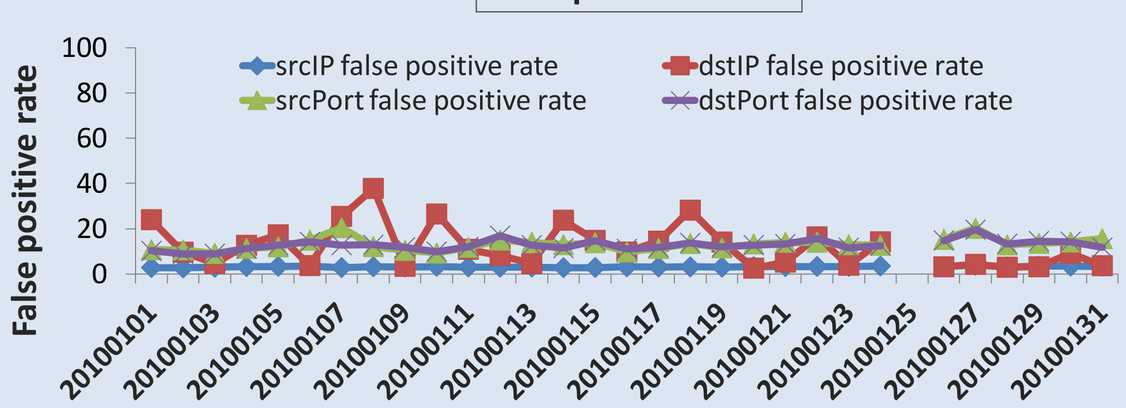

Fig. 4 False positive rate of detecting anomalous source IP, destination IP, source port, and destination port in traces collected on January 2010

➢ **Evaluation Dataset**: 30 backbone traffic traces from MAWI dataset [1] collected on January 2010 (~ 500,000 distinct IP/trace)
➢ **Results**: above 60% accuracy and 3-12% false positive rates (on average)

[1] K. Cho, K. Mitsuya and A. Kato. "*Traffic Data Repository at the WIDE Project*", USENIX 2000. Available at http://mawi.wide.ad.jp.
[2] R. Fontugne, P. Borgnat, P. Abry, K. Fukuda. "*MAWILab: Combining diverse anomaly detectors for automated anomaly labeling and performance benchmarking*", ACM CoNEXT 2010. Available at www.fukuda-lab.org/mawilab.

[1] Accuracy rate is computed by the number of anomalies that were correctly detected by our algorithm divided by the total number of anomalies that were detected by MAWILab [2]
[2] False positive rate is the total number of normal instances that were incorrectly detected as anomalies by our algorithm divided by the total number of normal instances in the trace.