Clustering Analysis for Malicious Network Traffic

Jie Wang, Lili Yang, Jie Wu and Jemal H. Abawajy

School of Information Science and Engineering, Central South University, Changsha, China

Email: jwang,liliyang@csu.edu.cn

Department of Computer and Information Sciences, College of Science and Technology, Temple University, Philadelphia, USA

Email: jiewu@temple.edu

School of Information Technology, Deakin University, Burwood 3125, Australia

Email: jemal@deakin.edu.au

Outline

- Introduction
- Malicious network traffic clustering scheme
- Clustering algorithm based on seeding-expanding
- Evaluation
- Conclusions

Introduction

- It is essential to identify what phase the attack is in as early as possible in order to detect attacks earlier.
- Clustering analysis is a good technology for classifying network traffic into different attack phases.
- However, because we cannot control the similarity level of data points in the clustering process, these cluster algorithms are not appropriate for detecting malicious network flows.
- New methods?

Outline

- Introduction
- Malicious Network Traffic Clustering Scheme
- Clustering Algorithm Based on Seeding-Expanding
- Evaluation
- Conclusions

Malicious Network Traffic Clustering Scheme

- Flow-level features

Name	Feature Discription
pkts	total number of packets
pkt-noPayload	total number of packets without payload
bytes	total number of bytes transferred
pay-bytes	total number bytes from all payloads
duration	flow duration
maxsz	maximum packet size
minsz	minimum packet size
avgsz	average packet size
stdsz	standard deviation of packet size
maxpy	maximum payload size
minpy	minimum payload size
avgpy	average payload size
stdpy	standard deviation of payload size
Flag	flags (acks, fins, resets, pushs, etc)

Malicious Network Traffic Clustering Scheme

- Feature preprocessing
  - A continuous-valued attribute is discretized by partitioning its range into multiple intervals.
  - Nominal attributes can be encoded using asymmetric binary attributes by creating a new binary attribute for each of the M states.
  - For an object with a given state value, the binary attribute representing that state is set to 1, while the remaining binary attributes are set to 0.

Malicious Network Traffic Clustering Scheme

- The process of malicious network traffic clustering from identifying flows includes:
  - Flow identification: flows are first constructed by aggregating traffic based on the 5-tuple flow identifiers
  - Feature extraction: features can be extracted from the flows
  - Feature preprocessing
  - Unsupervised learning and clustering

Outline

- Introduction
- Malicious Network Traffic Clustering Scheme
- Clustering Algorithm Based on Seeding-Expanding
- Evaluation
- Conclusions

Clustering Algorithm Based on Seeding-Expanding

- Similarity computing
  - The asymmetric binary similarity between the objects *i* and *j* can be computed as

$$sim(i,j) = \frac{q}{q+r+s}.$$

  - where *q* is the number of positive matches that equal 1 for both objects *i* and *j*, *r* is the number of attributes that equal 1 for object *i* but equal 0 for object *j*, and *s* is the number of attributes that equal 0 for object *i* but equal 1 for object *j*

Clustering Algorithm Based on Seeding-Expanding

- The SE algorithm contains the following steps:
  - Weight computing. For each data point d_i , the weight of d_i, $Weightd_i$, is the sum of $sim(d_i, d_j)$.
  - Seed selection. Firstly, all data points are sorted based on their Weights by decreasing and constructing a candidate queue $Q = \{q_1, q_2, \ldots, q_n\}$. Then, two data points, s_1 and s_2, are selected from Q as seeds. The first seed s_1 selected is as follows:

$$s_1 = \arg\max_{q_i} |Weight_{q_i} - Weight_{q_{i+1}}|$$

The second seed s_2 is the next of s_1 in the queue Q.

Clustering Algorithm Based on Seeding-Expanding

- Seed expanding. One data point, q, in the candidate queue is added to a cluster only if it has a maximal similarity value with seed s.
- Noise removal. Clusters with sizes smaller than 3 are considered noise data.

Outline

- Introduction
- Malicious Network Traffic Clustering Scheme
- Clustering Algorithm Based on Seeding-Expanding
- Evaluation
- Conclusions

Evaluation

- SE is evaluated by adopting a DDoS sample, which includes 5 stages:
  - IP-sweep of the AFB from a remote site;
  - Probe of the live IPs to look for the sadmind daemon running on Solaris hosts;
  - Breakings via the sadmind vulnerability, both successful and unsuccessful on those hosts;
  - Installation of the trojan mstream DDoS software on three hosts at the AFB;
  - Launching the DDoS.

Evaluation

- Results and discussion
  - There are three common indexes for evaluating the clustering results (purity, RI, F-Measure).
  - The results are illustrated in Figure 1, Figure 2, and Figure 3.
  - In these figures, the horizontal axis stands for a threshold, *r*, adopted by the SE algorithm. With the increase of *r*, cluster purity, the Rand Index, and the F-measure will increase.

Evaluation

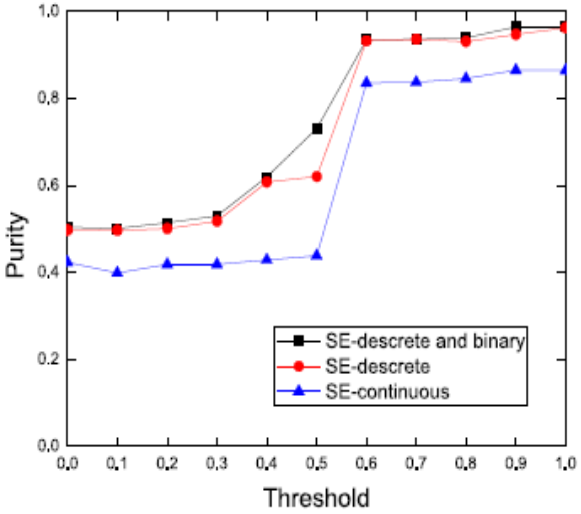

Fig. 1: Purity (SE)

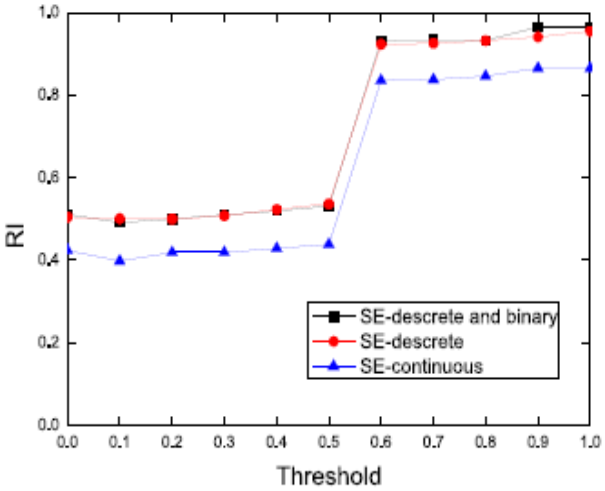

Fig. 2: Rand Index (SE)

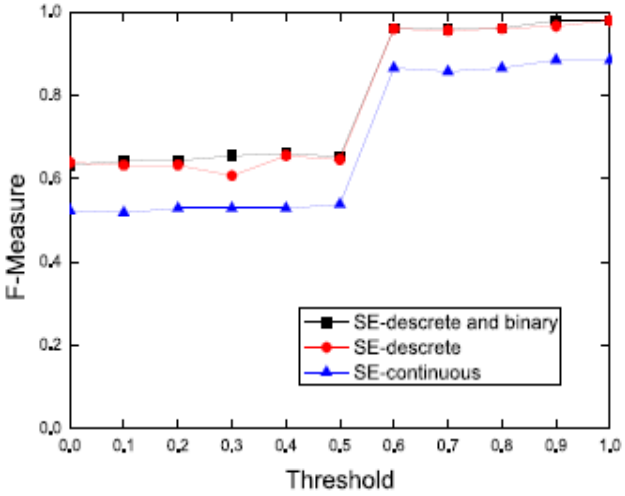

Fig. 3: F-measure (SE)

Evaluation

- Moreover, we compare the SE algorithm with K-Means, where we choose threshold $r = 0.9$. The results are shown in Table II.
- According to Table II, the SE algorithm can achieve a higher clustering performance than that of K-Means with any input cluster number.

TABLE II: The result comparing between SE and K-Means

	input cluster number	Purity	Rand Index	F-measure
SE-discrete		0.9552	0.9651	0.980306
K-Means-discrete	1	0.923751	0.854832	0.921735
K-Means-discrete	2	0.925504	0.920917	0.953059
K-Means-discrete	3	0.903593	0.92204	0.953408
K-Means-discrete	4	0.909728	0.923537	0.954257
K-Means-discrete	5	0.923751	0.925695	0.955486
K-Means-discrete	6	0.928133	0.926251	0.955806
K-Means-discrete	7	0.928133	0.94195	0.964883
K-Means-discrete	8	0.935145	0.942144	0.964992

Evaluation

- In the next experiments, we choose one seed, two seeds, three seeds, four seeds, five seeds, and six seeds in the process of seed selection, and the clustering results are illustrated in Fig.4, Fig.5, and Fig.6.

Evaluation

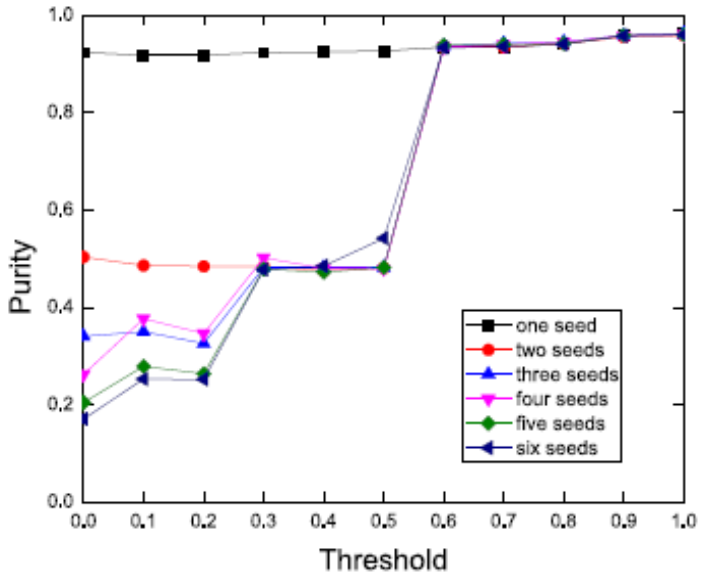

Fig. 4: Purity (SE) with different seed numbers

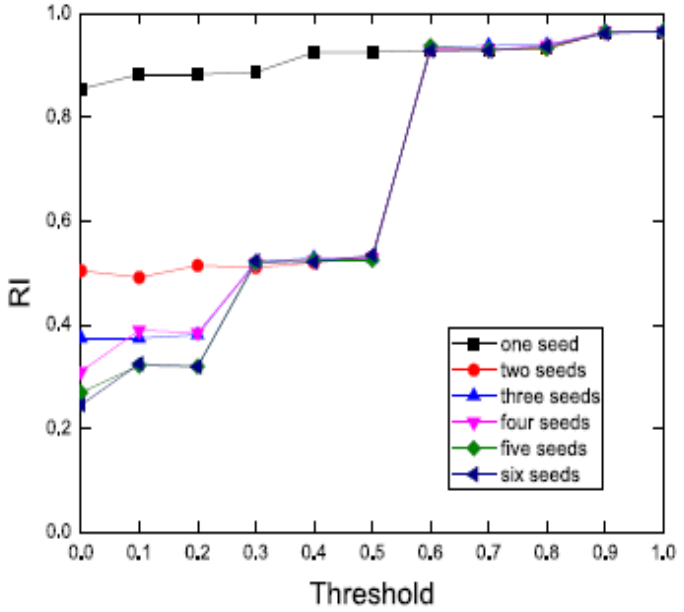

Fig. 5: Rand (SE) with different seed numbers

Evaluation

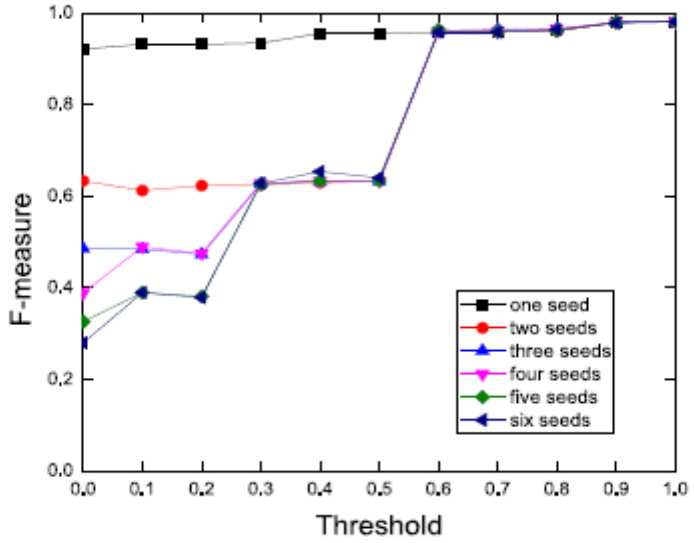

Fig. 6: F-measure (SE) with different seed numbers

Figure 4 illustrates the purity values with different numbers of clustering seeds. It is easy to find that clustering can obtain good results no matter how many seeds are selected for each iteration when threshold *r* is larger than 0.5. A similar conclusion can be drawn from Figures 5 and 6.

Evaluation

- We compare the number of iterations required with different seed numbers. The results of the comparison are illustrated in Figure 7.
- According to Figure 7, when the number of seeds increases, the number of iterations decreases gradually. However, when the number of seeds further increases, the iteration number remains the same.

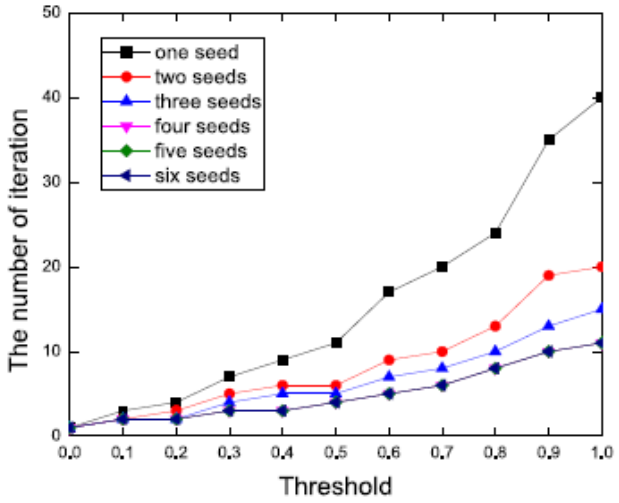

Fig. 7: The number of iterations required with different seed numbers

Outline

- Introduction
- Malicious Network Traffic Clustering Scheme
- Clustering Algorithm Based on Seeding-Expanding
- Evaluation
- Conclusions

Conclusions

- In this paper, we propose a clustering scheme based on Two-Seed-Expanding, which clusters attack flows into different phases.
- We also discuss the clustering's effectiveness when different seed numbers are used in the Seed-Expanding algorithm.
- Experimental results show Two-Seed-Expanding is better than K-Means and other kinds of Seed-Expanding that use different seed numbers in attack flow clustering.
- Next, we will do further research on how to identify attack flows from normal flows based on the clustering results.

Thank you!

Comments and questions are welcome!

Central South University