Poster: A DDoS Peak Traffic Volume Prediction Method Based on Machine Learning

Shuang Wei†, Yijing Ding†, Tongxin Li†, Shuaifu Dai‡, Xinfeng Wu† and Xinhui Han†*
†Peking University ‡CNCERT/CC
{shuangwei, dingyijing, litongxin, wuxinfeng, hanxinhui}@pku.edu.cn, daishuaifu@chanct.com

***Abstract*—DDoS defense nowadays relies on expensive and proprietary hardware appliances. When a massive attack begins, improper choices such as choosing fewer appliances or those without enough capacity may lead to more severe damage. As the previous work proposed[1], the choice heavily depends on the peak volume of the attack traffic(measured by packets per second). However, no prediction methods have been proposed to the best of our knowledge. In this paper we propose a method called *DDoSPVPredictor* to predict the peak volume of the DDoS traffic both effectively and efficiently. Based on machine learning, *DDoSPVPredictor* can predict the peak with only 24 features and for each attack the procedure can be finished in about 1.2s.**

We evaluate our solution's prediction accuracy using the 1998 MIT DARPA dataset. Result shows that *DDoSPVPredictor* is able to predict the peak volume of attack traffic with an accuracy of 85%. Therefore *DDoSPVPredictor* can help a lot in defending against massive DDoS attacks by optimizing its mitigation method using the predicted outcome.

I. Introduction

Distributed denial-of-service (DDoS) is one of the greatest threats to the Internet nowadays. In order to launch such attacks, the adversary takes control of many infected hosts and uses these hosts to flood the victim either by consuming the bandwidth or the resources of the victim. At Oct. 17th, 2016 7:00 a.m. EST, the DNS infrastructure managed by Dyn experienced a two hours' DDoS attack[2]. Many marquee brands such as Twitter and GitHub, were affected and couldn't be reached. This incident shows that large-scale DDoS attacks are very difficult to be defended against although they are easily detected.

For now, all DDoS defense methods use expensive and proprietary hardware appliances[3]. When a DDoS attack begins, making decisions on whether to absorb the traffic or shift the traffic to other appliances and which appliance to choose is quite a complicated issue. Improper policies may lead to longer downtime or a waste of resources. Moura[1] stated that the best choice depends on the peak volume of the attack traffic. However, there is no such method proposed. In this paper we proposed a prediction method called *DDoSPVPredictor* to predict the peak volume of the traffic at an early time during the attack.

II. Design and Implementation

In this section we describe the architecture of the *DDoSPVPredictor* and elaborate on the features used in prediction module.

* Corresponding Author.

A. Architecture

Fig.1 shows the basic architecture of *DDoSPVPredictor*. It consists of 2 main parts: `Attack Detector` and `Traffic Predictor`. First, `Attack Detector` detects the attack and inputs the traffic to `Traffic Predictor`. Then `Traffic Predictor` extracts features from traffic data and predicts the future peak traffic volume of the attack. The model we use is trained by the `History DDoS Dataset` before the attack.

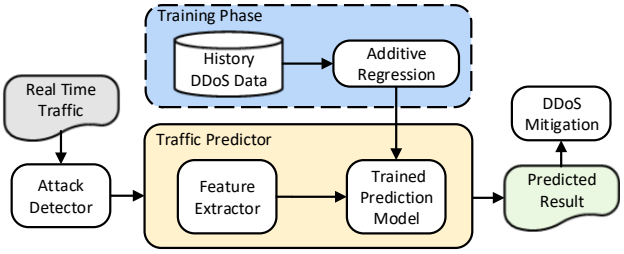

Fig. 1. Architecture of *DDoSPVPredictor*

1) Attack Detector: We apply Shiaeles' detection method[4] in this component. We call the time when the attack is detected T_d and the estimated time when attack begins T_s. We use the 10-minute-traffic before the attack as normal traffic and the traffic between T_s and T_d as attack traffic. Once the attack is detected, we input both the normal traffic and the attack traffic into `Traffic Predictor`.

2) Traffic Predictor: `Traffic Predictor` consists of two components, `Feature Extractor`, and `Trained Prediction Model`. When a DDoS attack is detected, `Feature Extractor` extracts the features representing traffic characteristics. Then put the features into `Trained Prediction Model` to predict the peak volume. After the DDoS we'll label the attack with its real peak traffic volume and store it in `History DDoS Dataset` for model training. By updating the dataset, the predictor adapts to new kinds of DDoS attacks constantly.

B. Feature selection

DDoSPVPredictor uses the following features for regression. We compute the following features, except the *Traffic.trend*, both from the normal traffic and attack traffic for comparison, and use all the calculated features for model training.

Srcip.entropy: It is used to describe the IP distribution, calculated by the IP list of the input traffic.

DstPort.entropy: It is used to describe port distribution. It can be used to distinguish the bandwidth consuming DDoS to service occupying DDoS.

Traffic.avg: It refers to the average packet number per second of the input traffic.

PktSize.avg: Mean of the packet size of the input traffic.

PktSize.entropy: Entropy of the packet size of the input.

Tcp&Udp.ratio: The ratio of the packets using TCP/UDP.

Icmp.ratio: The ratio of the packets using ICMP. Combined with the *Tcp&Udp.ratio*, it can be used to differentiate varieties DDoS types.

Traffic.trend: It consists of 10 features. We divide the traffic between T_s and T_d into 10 parts based on the packet timestamp. This results in a 10 dimensional row vector, from $trend_1$ to $trend_{10}$. The *ith* dimension represents the packet number in the *ith* time interval. We use these features to represent the trend of the traffic.

III. Experiments and Results

In Shiaeles' method, which we choose for `Attack Detector`, there is a strong linear relation between the number of packets and analysis time. As a 20000 packets' dataset DDoS event is detected in 2.4s, we assume DDoS attacks in our dataset, whose average packet number is around 70000, the event can be detected in 8.4s. So we assume the time interval between T_s and T_d is 8.4s.

For `Traffic Predictor`, we use SVR(SVM Regression)[5]. SVR is now the first choice for non-stationary series forecasting, because of its good generalization ability and guaranteeing global minima. For the prediction problems, the relation between the features and the predicted results is non-linear, a mapping methods need to be applied. So we build our prediction model using polynomial kernel as the kernel function. We use MIT DARPA 1998 training dataset, from which we find 127 valid DDoS incidents. Our prediction model was implemented using Weka 3-7.

A. Comparison with Other Methods

As a comparison, we do several experiments on different machine learning algorithms, such as Linear Regression, Random Forest. Results show that SVR performs best and can lead to the minimum error rate, as can been seen in Fig. 2.

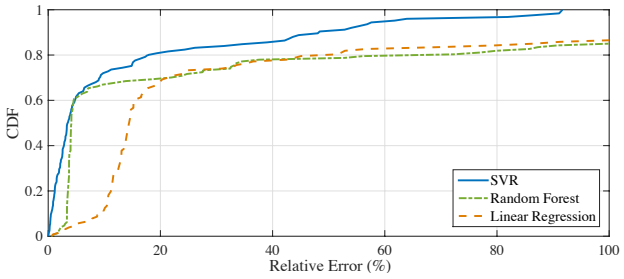

Fig. 2. Comparison results of different training methods

B. Feature Validation

We divide the features into 4 groups: **Addr.** group includes *Srcip.entropy* and *DstPort.entropy* which are referred to the address information such as IP, port. **Traf.** group are composed of *Traffic.avg* and *Traffic.trend* which are all related to traffic volume. **PktSz.** group includes *PktSize.avg* and *PktSize.entropy*. **Prto.** are composed of *Tcp/Udp.ratio* and *Icmp.ratio*. In order to validate the impact of different groups of features, we conduct experiments with different combinations by removing each group of features respectively. Table I shows the results, with w/o denotes experiments without the corresponding group of features and ARE denotes average relative error. It shows that all the features play indispensable roles in our method.

TABLE I
The Impact of Different Features Groups to ARE

Method	All	w/o Addr.	w/o Traf.	w/o PktSz.	w/o Prto.
ARE(%)	15.87	16.61	36.62	19.11	18.25

C. Experimental Results

It takes our predictor 30s to extract features from 127 incidents. The average time spent on each incident is 0.2s. Time taken to do the prediction is less than 1s. So for one attack we can finish the prediction in 1.2s. We use 10-fold cross-validation to evaluate our predicted results. The results show that our method can predict the peak traffic volume with an ARE of 15.87%. And according to Fig.2, for 80% of the test cases we can limit the prediction error into 20%.

IV. Conclusion

This paper proposed a novel prediction method called *DDoSPVPredictor* to forecast the peak volume of the DDoS attack traffic at an early time. With the help of our predicted results, the DDoS defense system can better decide which mitigation plan should be carried out for each attack. We build our prediction model using SVR. Then we test its prediction accuracy on MIT DARPA 1998 training dataset using 10-fold cross-validation evaluation method. Results show that *DDoSPVPredictor* limits the average relative error in 15.87%, which is precise enough to provide a reference for DDoS defense policy choice.

Acknowledgment

This paper is supported by the National Science Foundation of China under Grant 61402125.

References

[1] G. C. Moura, R. d. O. Schmidt, J. Heidemann, W. B. de Vries, M. Müller, L. Wei, and C. Hesselman, "Anycast vs. ddos: Evaluating the november 2015 root dns," in *Proceedings of the ACM Internet Measurement Conference (IMC 2016). Santa Monica, CA, USA. November*, 2016.

[2] K. York, "Dyn statement on 10/21/2016 ddos attack," http://dyn.com/blog/dyn-statement-on-10212016-ddos-attack/, October 2016, accessed on 08/01/2017.

[3] S. K. Fayaz, Y. Tobioka, V. Sekar, and M. Bailey, "Bohatei: flexible and elastic ddos defense," in *Usenix Conference on Security Symposium*, 2015.

[4] S. N. Shiaeles, V. Katos, A. S. Karakos, and B. K. Papadopoulos, "Real time ddos detection using fuzzy estimators," *computers & security*, vol. 31, no. 6, pp. 782–790, 2012.
[5] C. C. Chang and C. J. Lin, "Libsvm: A library for support vector machines," *Acm Transactions on Intelligent Systems & Technology*, vol. 2, no. 3, p. 27, 2011.