Analyzing Denial-of-service Attacks in KDD CUP 99 Data Set for Intrusion Detection System

Kyaw Thet Khaing, Thinn Thu Naing
University of Computer Studies, Yangon
Kyawthetkhaing.ucsy@gmail.com , thinnthu@gmail.com

Abstract

Recently cyber security has emerged as an established discipline for computer systems and infrastructures with a focus on protection of valuable information stored on those systems from adversaries who want to obtain, corrupt, damage, destroy or prohibit access to it. One of the security-threat which is difficult to address using traditional network security techniques is Denial of Service (DoS) attacks. We have seen increasing numbers of denial of service (DoS) attacks against online services and web applications either for extortion reasons, or for impairing and even disabling the competition. This paper implemented that how much these (DoS) attacks stand on a top role above on the other attacks. We evaluate those results on four attack categories as found in the KDD Cup 99 intrusion detection datasets, which is widely used as one of the publicly available data sets for network-based anomaly detection system such as Intrusion Detection System (IDS).

1. Introduction

According to the Wikipedia, the denial-of-service attack is an attempt to make a computer resource unavailable to its intended users. Although the means to carry out, motives for, and targets of a DoS attack may vary, it generally consist the concerted efforts of a person or people to prevent an internet sites or service from functioning efficiently or at all, temporarily or indefinitely. Perpetrators of DoS attacks typically target sites or services hosted on high-profile web servers such as banks, credit card, payment gateways, and even root name servers.

On common method of attack involves saturating the target (victim) machine with external communications requests, such that it cannot respond to legitimate traffic, or responds so slowly as to be rendered effectively unavailable. In general terms, DoS attacks are implemented by either forcing the targeted computer(s) to reset, or consuming its resources so that it can no longer provide its intended service or obstructing the communication media between the intended users and the victim so that they can no longer communicate adequately.

Denial-of-service attacks are considered violations of the IAB's Internet proper use policy, and also violate the acceptable use policies of virtually all Internet Service Providers (ISP). They also commonly constitute violations of the laws of individual nations[1].

We studied the details of the research done in attacks on KDD Cup 99 intrusion detection data sets, which is widely used as one of the few publicly available data sets for network-based anomaly detection system such as intrusion detection system. These data sets are public by at Massachusetts Institute of Technology (MIT) Lincoln Lab [3].

The first important deficiency in the KDD dataset is the huge number of network traffic records. In "kddcup_data_corrected" dataset, it consists 4898431 connections record and take 721MB in file size. This large amount of records in data set will the evaluation results to be biased by the methods which have better detection rates on these records.

In addition, to analyze the difficulty level of the record in KDD data set, we use "kddcup_10_ percent_corrected" dataset with only 494021 in record, which are only ten percent of the entire data set for dataset1. In the first dataset, about 78% of the records are duplicated [2]. These redundant records will cause learning algorithms to be biased towards the more frequent records, and thus prevent it from learning unfrequented records which are usually more harmful to networks such as U2R attacks. So, we used "corrected" data set for dataset2 and "kddtest+" dataset for dataset3 which is a new-version of KDD dataset [2, 4]. These dataset2 and dataset3 contain 292300 and 125973 respectively.

The rest of the paper is organized as follows. Section II introduces the KDDCUP99 data set which widely used in anomaly detection. In Section III, we

about the Intrusion Detection System were described. Details analysis of DoS attacks with its several attack types was implemented in Section VI. We concluded our paper and future extension in Section V.

2. KDD CUP 99 Data Set Description

Since 1999, KDD'99 [3] has been the most widely used data set for the evaluation of anomaly detection methods. This data set is built based on the data captured in DARPA'98 IDS evaluation program [5]. DARPA'98 is about 4 gigabytes of compressed raw (binary) tcpdump data of 7 weeks of network traffic, which can be processed into about 5 million connection records, each with about 100 bytes. The two weeks of test data have around 2 million connection records. KDD training dataset consists of approximately 4,900,000 single connection vectors each of which contains 41 features and is labeled as either normal or an attack, with exactly one specific attack type. The simulated attacks fall in one of the following four categories:

(1) Denial of Service Attack (DoS): is an attack in which the attacker makes some computing or memory resource too busy or too full to handle legitimate requests, or denies legitimate users access to a machine.

(2) User to Root Attack (U2R): is a class of exploit in which the attacker starts out with access to a normal user account on the system (perhaps gained by sniffing passwords, a dictionary attack, or social engineering) and is able to exploit some vulnerability to gain root access to the system.

(3) Remote to Local Attack (R2L): occurs when an attacker who has the ability to send packets to a machine over a network but who does not have an account on that machine exploits some vulnerability to gain local access as a user of that machine.

(4) Probing Attack: is an attempt to gather information about a network of computers for the apparent purpose of circumventing its security controls.

It is important to note that the test data is not from the same probability distribution as the training data, and it includes specific attack types not in the training data which make the task more realistic. Some intrusion experts believe that most novel attacks are variants of known attacks and the signature of known attacks can be sufficient to catch novel variants. The datasets contain a total number of 24 training attack types, with an additional 14 types in the test data only. The name and detail description of the training attack types are listed in [7].

KDD'99 features can be classified into three groups:

(1) Basic features: this category encapsulates all the attributes that can be extracted from a TCP/IP connection. Most of these features leading to an implicit delay in detection.

(2) Traffic features: this category includes features that are computed with respect to a window interval and is divided into two groups:

(a) "same host" features: examine only the connections in the past 2 seconds that have the same destination host as the current connection, and calculate statistics related to protocol behavior, service, etc.

(b) "same service" features: examine only the connections in the past 2 seconds that have the same service as the current connection.

The two aforementioned types of "traffic" features are called time-based. However, there are several slow probing attacks that scan the hosts (or ports) using a much larger time interval than 2 seconds, for example, one in every minute. As a result, these attacks do not produce intrusion patterns with a time window of 2 seconds. To solve this problem, the "same host" and "same service" features are re-calculated but based on the connection window of 100 connections rather than a time window of 2 seconds. These features are called connection-based traffic features.

(3) Content features: unlike most of the DoS and Probing attacks, the R2L and U2R attacks don't have any intrusion frequent sequential patterns. This is because the DoS and Probing attacks involve many connections to some host(s) in a very short period of time; however the R2L and U2R attacks are embedded in the data portions of the packets, and normally involves only a single connection. To detect these kinds of attacks, we need some features to be able to look for suspicious behavior in the data portion, e.g., number of failed login attempts. These features are called content features.

3. Overview of Intrusion Detection System

3.1 Intrusion

Intrusions are actions that attempt to bypass security mechanisms of computer systems. So, they are any set of actions that threatens the integrity,

availability, or confidentiality of a network resource. These properties have the following explanations:

Confidentiality – means that information is not made available or disclosed to unauthorized individuals, entities or processes;

Integrity – means that data has not been altered or destroyed in an unauthorized manner.

Availability – means that a system or a system resource that ensures that it is accessible and usable upon demand by an authorized system user. [9]

3.2 Intrusion Detection System

An Intrusion detection system is used to detect several of malicious behaviors that can compromise the security and trust of a computer system. This includes network attacks against vulnerable services, data driven attacks on applications, host based attacks such as privilege escalation, unauthorized logins and access to sensitive files, and malware (virus, Trojan, horses, and worms).

The technologies of intrusion detection system are indispensable for network and computer security, as the increasing the serious matters cause by cyber threats. Intrusion Detection is the process of detecting these cyber attacks in a system or network by monitoring. Intrusion Detection System (IDS) monitors network traffic for untrusting activity and warning the system or network administrator against malicious attacks. The goal of Intrusion Detection System is to alert and protect the confidentiality, integrity and availability of critical networked information systems.

There can be divided into two main approaches named misuse and anomaly detection. Misuse detection is based on a description of known malicious activities. This description is often modeled as a set of rules referred to as attack signatures. An anomaly detection IDS looks for anomalies, meaning it thinks outside of the ordinary. It uses rules or predefined concepts about "normal" and "abnormal" system activity (called heuristics) to distinguish anomalies from normal system behavior and to monitor report on, or block anomalies as they occur [7].

An intrusion detection system (IDS) is a software and or hardware designed to detect unwanted attempts at accessing, manipulating, and/or disabling of computer systems, mainly through a network, such as the Internet. These attempts may take the form of attacks, as examples, by crackers, malware and/or disgruntled employees. IDS cannot directly detect attacks within properly encrypted traffic.

An IDS can be composed of several components. Sensors which generate security events, a Console to monitor events and alerts and control the sensors, and a central Engine that records events logged by the sensor in a databases and uses a system of rules to generate alerts from security events received. These are several ways to categorize an IDS depending on the type and location of the sensors and the methodology used by the engine to generate alerts. In many simple IDS implementations all three components are combined in a single device or appliance [8].

4. Detail analysis on DoS

Denial-of-service (DoS) attacks cost business millions of dollars each year because of system downtime, lost revenue and productivity, tarnished reputation, and the hours required by technical staff to locate the problem and resolve it. Once customers lost confidence in the security of the systems holding their confidential and financial information, they will often take their business elsewhere.

Classification of the types of DoS attack is also important because since the different types of DoS attacks employ slightly different attack mechanism, this means that the defense against them is also different.

As the formerly express, KDD data sets classified Denial-of-service DoS in 6 attack types such as smurf, teardrop, neptune, land, pod, and back. Among them smurf and neptune have a large amount of record in three datasets.

Smurf attack is a type of network-level by overwhelming the victim machine with Internet Control Message Protocol (ICMP) echo replies from computers in the same broadcast network computers in the same broadcast network by sending forged ICMP echo request to an IP broadcast address using the IP address of the victim machine, making computers in the same reply to the requests, flooding the victim machine with ICMP echo replies.

The important role of classification on DoS types are shown on table 1. We can simply seen that the smurf and neptune attacks among other types are significantly take large amount for attacking us.

Table 1. Matrix of comparison among different types in DoS attack

Attack Types	Data Set 1	Data Set 2	Data Set 3
Smurf	280790	164091	2646
back	2203	1098	956
Land	21	9	18
Pod	264	87	201
Teardrop	979	12	892
Neptune	107201	58001	41214

The number of connection records that represented for DoS attacks of data sets is shown in

Table 2. And Figure 1 and 2 had shown the comparison the amount of attacking in data sets among DoS attacks and other attacks according to the number of connection records in datasets.

Table 1. Number of network connection in three data sets

no. of Connection	Dataset 1	Dataset 2	Dataset 3
Total	494021	292300	125973
DoS	391458	223258	45927
U2R	52	39	52
R2L	1126	5993	995
Probe	4107	2377	11656
normal	97278	60593	67343

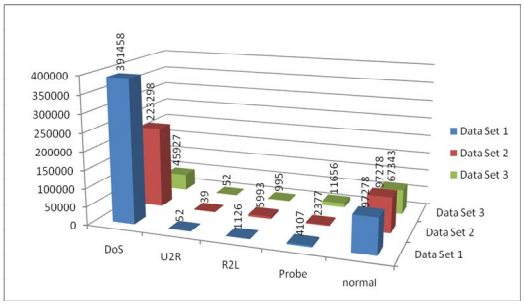

Figure 1. Comparison of attack rate on three datasets

We can clearly seen that the number of connection of DoS attacks in all datasets significantly overwholm the network traffic even normal state connection can not take 25% of the traffic in first two datasets. In this state, the redundant records play a role on dataset3. The DoS attacks fall down under normal traffic can seen definitely. However, according to the Figure 2, network traffic porbability have to suffer in DoS attacks 86.81% on overall network traffic.

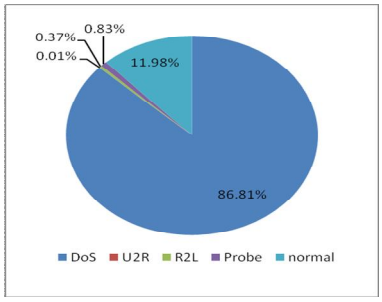

Figure 2. Overall percentages of attacks hold on three data sets

5. Conclusion

We have seen increasing numbers of denial of service (DoS) attacks against online services and web applications either for extortion reasons, or for impairing and even disabling the competition. After analyzing above three KDD CUP 99 data sets, the results show that DOS attacks is the most highly possible attacks and the effect of DOS attacks is hazard in every system. So this attack should be viewed as a risk management issue that can be effectively dealt with like other business. This means minimizing exposure where possible and being prepared should an attack eventuate.

6. References

[1] "Denial-of-service attack", 2009. Available on: http://en.wikipedia.org/ wiki/denial-of-service_attack

[2] M. Tavallace, E. Bagheri, W. Lu, and A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set", *Submitted to second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CSIDA)*, 2009.

[3] KDD Cup 1999. Available on: http://kdd.ics.uci.edu/ databases/kddcup99/ kddcup99.html, December 2009.

[4] "NSL-kdd data set for network-based intrusion detection system." Available on: http://nsl.cs.unb.ca/NSL-KDD/ , December 2009.

[5] KDD Cup DARPA 1998. Available on: http://kdd.ics.uci.edu/databases/kddcup98/ kddcup98.html, December 2009.

[6] MIT Lincon Labs. 1998 DARPA Intrusion Detection Evaluation. Available on: http://www.ll.mit.edu/mission/ communications/ist/corpra/idvel/index.html. December,2009.

[7] M. Bahrololum, E. Salahi and M. Khaleghi, "Anomaly Intrusion Detection Design using Hybrid of Unsupervised and Supervised Neural Network", *International Journal of Computer Network & Communications(IJCNC), Vol.1, No.2*, July 2009.

[8] "Intrusion detection system", 2009 http://en.wikipedia.org/ wiki/intrusion_detection_sytem

[9] V. Marinova-Boncheva, "A Short Survey of Intrusion Detection System", 2007.