

DDoS Detection and Mitigation Using Machine Learning

Arpit Gawande

March 21, 2018

A thesis submitted to the Graduate School - Camden
Rutgers, The State University of New Jersey

for the degree of Master of Science

Graduate Program in Scientific Computing

Written under the guidance of
Dr. Jean-Camille Birget

Contents

1	Existing Systems	3
1.1	What it DDoS attack	3
1.2	Challenges in dealing with DDoS attacks	3
2	DDoS Detection and Mitigation	5
3	Network Functioning	5
3.1	Point of knowledge	5
4	Our Approach	6
5	Implementation	8
5.1	Machine Learning	12
5.1.1	Feature Scaling	13
5.1.2	Clustering	13
5.1.3	Anomaly Detection using One Class Support-Vector Machine	18
6	Detection	20
7	Conclusion	22

Abstract

Distributed Denial of Service (DDoS) attacks are very common these days [1]. So it is evident that current industry solutions such as completely relying on Internet Service Provider (ISP) or setting up DDoS defense infrastructure are not sufficient in detecting and mitigating DDoS attacks, hence consistent research is needed. Most of the current industry solutions involve setting up centralized expensive hardware system which can analyze the packets¹ [2] for probable DDoS attacks. Also each router provider has different protocols to communicate the between the DDoS attack detection system and the router limiting the reach of DDoS detection systems. In this paper we are going to discuss a way to detect DDoS attack using machine learning tools at the routers, also we will propose a communication architecture for mitigating it.

1 Existing Systems

1.1 What it DDoS attack

Distributed Denial of Service (DDoS) attack is the way to jam host network or its resources with large number of data packets or connection, so that host become disabled to serve. There are different types of DDoS attacks such as 1. Volume based e.g. SYN Flood Attacks, in which victim is flooded with hight volume of packets or connection. 2. Application based, in which application such as DNS, VOIP or HTTP where attacked. 3. Low rate DDoS attacks, in which attacker exploit the vulnerability in application design, e.g. Slowloris. [3]

1.2 Challenges in dealing with DDoS attacks

The real challenge in detecting and defending DDoS attack is because of its dynamic nature. The source² is not a single node or system on the Internet but there can be many systems participating in DDoS attack, and often these systems are distributed over different regions of the Internet. Also the source of the packet is often spoofed³ [4], which makes harder to know the actual IP address of system from where attack is originated because original attack source is changed in spoofed data packet. On top of that, many times the source system itself is not aware that it is compromised and it is being used as a bot [5] by an attacker to launch DDoS attack.

Detecting and mitigating attack at the destination⁴ is not very useful as because destination

¹Messages that are sent on Internet are broken into shorter messages for transmission. These short messages are called packets. Term coined by Donald Watts Davies.

²It is a system/device on the Internet which has an IP address and which is involved in DDoS attack

³spoofing is the way to change the source IP address of the message. This is a known issue in the protocol iteself not in the implementation

⁴System under DDoS attack

may know that the attack is happening but to stop it happening it will have to block all the incoming traffic including the legitimate traffic, because source address can not be reliable way to know the attack source. To avoid this, many network device producing companies such as Cisco, Netgear have come up with some solutions. Many of the solutions provided by those giant or the research that is done in this field has been focusing on collecting network traffic flow information [6] (we will call it just flow) at routers(gateways) and then send that information to the central system for analysis. Central system is a hardware and software infrastructure which is capable of processing and analyzing large flow information.

Some of the major protocols which are widely used for flow collection and analysis are, Internet Protocol Flow Information Export (IPFIX) protocol created by Internet Engineering Task Force (IETF), Ciscos NetFlow [7] and Sflow(Sampled flow) [8]. These protocols have defined standard way to export flow information from router and similar devices. All these flow monitoring protocols gather information and send the consolidated flow information to the centralized server where user can login and perform functions; such as Security Monitoring, Bandwidth monitoring, Resource Management, Traffic Analysis, Performance Management etc. It will have some modules which are specifically used for anomaly/DDoS detection.

E.g. Cisco netflow has flow exporter, collector and analysis modules. Flow exporter modules are installed on routers. The routers which are having flow exporter modules, send flow information to collector module installed on the server. Along with the collector module server also has analysis module which can be used to detect different patterns in the flow.

These technologies scales well and can be sufficient to indicate trends in network traffic but they have limitations. 1) They are not cross platform, e.g. router with Sflow protocol can not work with Cisco routers. 2) They involve setting up expensive hardware which acts as collector server. 3) Source address is used for flow analysis which is not reliable due to IP spoofing in the case of of DDoS attack.

Now we know that router based flow analysis can be useful for anomaly detection but it has limitations. We don't want to set up expensive hardware, we want to have protocol or system which is compatible with other routers. Also we dont want source IP address for detection analysis. So if we can come up with the way by which we can detect anomaly in the traffic at the routers and create a communication protocol between the routers and the destination server or network, then routers and network can take better decisions on regulating the packet flow. If we use only the destination IP address for analysis, then we will be more efficient in detecting and mitigating DDoS attacks.

2 DDoS Detection and Mitigation

DDoS attacks can be detected by checking if there is any anomalous behavior in the network traffic, such as sudden increase in the number of packets going to a destination. This can be done at server by observing all the incoming traffic or it can be done by observing all the outgoing traffic at ISP or at every router. Attack can be mitigated if the anomalous packets are blocked reaching their destination.

3 Network Functioning

3.1 Point of knowledge

A switch creates a network and router connects networks. A router links computers to the Internet through other routers. Routers are the backbone of the network who helps to forward packets from one point to another point on the Internet. Every packet traveling on the Internet has to go through a router [9]. A router knows where the packet is destined hence it could serve as a first point of knowledge about the change in the flow information for a destination network. Each router has interfaces to which hosts or other networks are connected. So a router is aware to whom it is connected. Routers use protocols to communicate and by that they gather knowledge about other networks or routers on the Internet. ICMP [10] is one of the most frequently used protocol by routers to communicate.

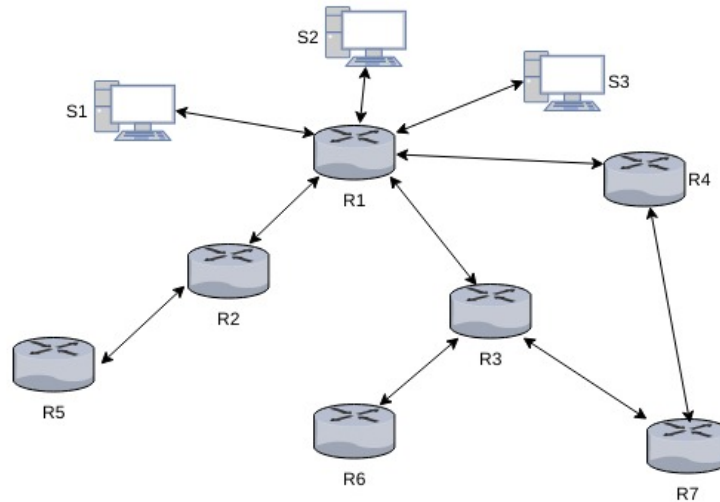


Figure 1: Network Example

Let's illustrate this using an example. In the above figure we can see that host S1, S2, S3 are connected to router R1. Router R1 is connected to Internet through router R2, R3, R4, thus

every packet reaching to S2 is coming from either of these three routers. All three routers are located in different geographical region. Most of the websites are regional, either county, state or national (If we leave out few global websites) and hence they are mostly accessed from those region it is meant for. E.g. Rutgers University website is accessed mostly from the eastern region of United States and that too mostly from the New Jersey State or the Philadelphia region.

Using traceroute we can find out how many hops⁵ away the destination is. Following is one of the captured traceroute for Rutgers University website.

```
arpit@omega:~$ traceroute camden.rutgers.edu
traceroute to camden.rutgers.edu (128.6.34.90), 30 hops max, 60 byte packets
 1 192.168.0.1 (192.168.0.1) 1.067 ms 1.697 ms 1.684 ms
 2 10.240.177.197 (10.240.177.197) 7.617 ms 9.975 ms 10.302 ms
 3 67.59.225.66 (67.59.225.66) 10.803 ms 12.759 ms 13.074 ms
 4 dstswr1-ge1-2.rh.mhwnj.cv.net (67.83.247.130) 18.962 ms 18.952 ms 18.902 ms
 5 67.59.239.121 (67.59.239.121) 18.844 ms 451be043.cst.lightpath.net (65.19.114.67) 18.314 ms
 6 451be031.cst.lightpath.net (65.19.98.49) 19.762 ms 64.15.3.138 (64.15.3.138) 10.763 ms 17.7
 7 * * *
 8 * * *
 9 RUTGERS-THE.ear3.Newark1.Level3.net (4.14.216.6) 33.338 ms 32.792 ms 33.274 ms
10 * * *
11 * * *
12 * * *
13 * * *
14 * * *
15 web-www.camden.rutgers.edu (128.6.34.90) 22.632 ms 23.859 ms 23.866 ms
```

Figure 2: Trace Route: All the routers in the path to destination

We can see that there are about 14 routers (if we don't consider the home router 192.168.0.1) to reach to the camden.rutgers.edu. This trace route is taken from a location in the New Jersey State.

4 Our Approach

From Figure 1 and 2 we know that routers are located at different geographical locations and also there is pattern in which the particular destination website is getting accessed from the different regions. Some of the service providers such as GeoIP or Google, can find out the location from where the traffic is coming in the network for a given destination, but that is approximate based on the source IP. In the case of DDoS attack this information is unreliable, because the packet source address is often spoofed. This makes difficult to know the actual geographical location from which the packet has come, but routers through which that packet has traveled can provide their own geographical information. Such information can be useful to understand the path through which packet has traveled and thus we can know the region from which the packet is originated.

In the normal scenario there is some definite pattern in which the website is accessed from different geographical regions, and this pattern can be learned over the time. Thus finding this

⁵hops are intermediate routers in the communication channel

pattern in the flow⁶ at the router can form the basis of analysis in this paper. Whenever there is a deviation from the normal flow pattern learned in the past for a particular destination then that change in pattern as well as router geographical region information can be communicated to the destination network. The destination network on receiving that information, can decide, whether it want the reporting router to discard or forward the traffic for it. This is a selective process in which traffic from only specific router is blocked while traffic from other routers remain unaffected

With the advance of electronic and the Internet of things, processing and storage capacity of the electronic devices has increased. Router are also not left behind, but storage capacity of router is always very less compared to storage server which collect flow data for network traffic analysis. If we use the learning techniques which don't need much storage then we don't have to store large chunk of packets on the router. Instead of storing data packets for longer time for analysis, we can learn from small number of packets and then discard packets once learning is done leaving behind only learned information on the router. This is necessary because the number of entries on the Internet routing table has steadily grown. Now that the table has passed 500,000 routes [11] so storing each and every flow information for these routes could be difficult.

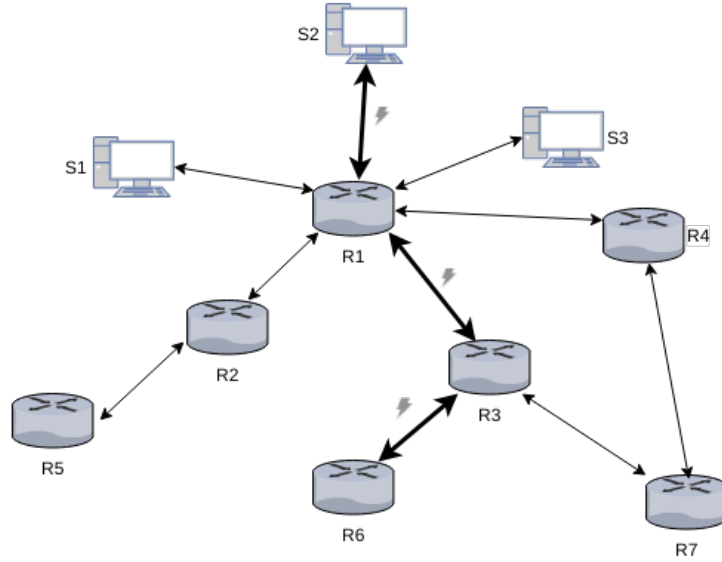


Figure 3: DDoS Attack path

In the above figure an attack is initiated from the region where router R6 is located and from router R6 data packets travel to the victim⁷ system S2. Attack packets traveled through router R3 and R1 to reach system S2. If we can detect attack at router R6, then router R6 can

⁶In this paper we will consider flow only in sense of destination address

⁷Victim is a computer system which is under DDoS attack

discard all the packets heading towards system S2. In this process, only traffic from router R6 is affected but traffic from all other routers remain unaffected.

To achieve this, we will gather the flow information during a time window (e.g 300 sec) whose size will be fixed at the beginning. These time windows can be combined to form a period which is a portion of a day during which traffic is measured. Once we have flow information we can apply learning techniques on each flow iteratively to gain deeper knowledge about normal behavior of the flow.

In the proposed system, each router will itself act as a analyzer. Each packet will be analyzed and flow statistic is created based on the destination IP address. Based on the statistic, clustering of destination IP address using input feature vector is done. Clustering is the process of examining a collection of points, and grouping the points into clusters according to some distance measure. The goal is to minimize the distance of the point in the cluster to other points in the same clusters. [12].

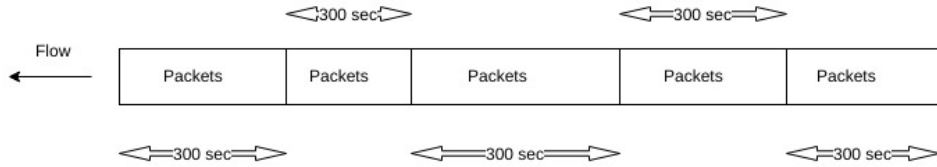


Figure 4: Data Flow Segments

Once the clustering is learned it will be used as bench mark for all future flow. Router will constantly keep clustering destination IP address and if there is deviation in the normal traffic at router for any destination then that will affect the clustering and it will cause destination to be placed in different cluster. This change in the cluster for a given distention can be marked as change in the behavior of the traffic for that destination. Along with the cluster we will also use Novelty Detection algorithm to achieve more accurate result. This change in traffic pattern will be reported to destination network, which then decide on regulating the traffic to itself on the router which has send the information

5 Implementation

There are different types of DDoS attacks, such as Volume based, Application based and also Low rate DDoS attacks. Among these different types of DDoS attacks, the volume based attacks are most common. In the volume based attack, victim is flooded with a high volume of Internet packets (TCP, UDP, HTTP or ICMP) make it unable to serve the requests.

For the demonstration of suggested approach we are simulating a volumed based Bot attack. Bots are the compromised computer systems, controlled by attacker for launching an attack. They are not bounded by geographical boundaries, so they can be anywhere in the Internet. Botnet(Network of bots) are employed by an attacker to launch a DDoS attack. As we know that the Internet is connection of different computer system that communicate with each other, through different channels such as cables, satellite or radio device and these communication channels run through out the glob; connecting different computer systems at different locations.

We are using Wireshark, an open source tool, for capturing Internet packets. Wireshark can capture all digital information received or send through different devices such as Ethernet or wifi devices, which connects computer to the Internet. It also helps identify different protocol packets (e.g. TCP, UDP) within the packet created at Data Link Layer packet. Data Link Layer packet is a wrapper over all higher protocols such as TCP/UDP in OSI model.

No.	Time	Source	Destination	Protocol	Length	Info
1	0.000000000	192.168.0.7	173.194.175.189	QUIC	107	CID: 10146815167217732570, Seq: 49
2	0.049933000	173.194.175.189	192.168.0.7	QUIC	152	CID: 0, Seq: 56
3	3.105773000	192.168.0.7	192.168.0.10	TCP	223	[TCP segment of a reassembled PDU]
4	3.107726000	192.168.0.10	192.168.0.7	TCP	263	[TCP segment of a reassembled PDU]
5	3.107974000	192.168.0.7	192.168.0.10	TCP	108	52718-8009 [ACK] Seq=116 Ack=118 Win=346 Len=0 TSval=32023603 TSecr=254463
6	4.701349000	192.168.0.7	52.204.61.141	TLSv1.2	396	Application Data
7	4.717465000	52.204.61.141	192.168.0.7	TCP	146	443-40122 [ACK] Seq=1 Ack=209 Win=314 Len=0 TSval=1050903052 TSecr=32024001
8	4.718005000	52.204.61.141	192.168.0.7	TLSv1.2	471	Application Data
9	4.718249000	192.168.0.7	52.204.61.141	TCP	108	40122-443 [ACK] Seq=289 Ack=326 Win=727 Len=0 TSval=32024005 TSecr=1050903052
10	7.255990000	192.168.0.7	54.89.16.99	TCP	108	40118-443 [ACK] Seq=1 Ack=1 Win=237 Len=0 TSval=32024640 TSecr=3880385116
11	8.001888000	65.52.108.76	192.168.0.7	TLSv1.2	1351	Application Data
12	8.013787000	192.168.0.7	65.52.108.76	TCP	1536	[TCP segment of a reassembled PDU]
13	8.014102000	192.168.0.7	65.52.108.76	TLSv1.2	1229	Application Data
14	8.035666000	65.52.108.76	192.168.0.7	TCP	146	443-39116 [ACK] Seq=1206 Ack=2550 Win=514 Len=0 TSval=119147688 TSecr=32024829
15	8.108997000	192.168.0.7	192.168.0.10	TCP	223	[TCP segment of a reassembled PDU]

Frame 1: 107 bytes on wire (856 bits), 107 bytes captured (856 bits) on interface 0
 ▶ Radiotap Header v0, Length 14
 ▶ IEEE 802.11 QoS Data, Flags: .p.....T
 ▶ Logical-Link Control
 ▶ Internet Protocol Version 4, Src: 192.168.0.7 (192.168.0.7), Dst: 173.194.175.189 (173.194.175.189)
 ▶ User Datagram Protocol, Src Port: 37254 (37254), Dst Port: 443 (443)
 ▶ QUIC (Quick UDP Internet Connections)

```

0000  00 00 00 00 00 00 00 00 00 00 07 04 07 08 41 .....A
0010  00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 .....S.....S
0020  00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 .....S.....S
0030  aa aa 03 00 00 00 00 00 45 00 00 33 c1 93 40 00 .....E..3..@
0040  40 11 5a f7 c0 a8 00 07 ad c2 af bd 91 95 01 bb @.Z.....
0050  00 1f 44 a9 0c da 00 8f 60 a2 ba 00 0c 31 d5 00 .....K.....l
0060  1f 06 73 54 c3 d5 ea 93 a9 02 7f .....sT.....
  
```

Figure 5: Wireshark Tool: snippet of captured packets

To gather data for the demonstration, we have created a small network which has one router and couple of host machine. Each machine can be a victim of DDoS attack. We have installed Wireshark tool on one of the machine in the this network. For capturing traffic in the network we are using the Promiscuous mode of Wireshark. In Promiscuous mode, network interface can record not only the traffic that is intended to itself but all of the traffic on the network, so we can see all in out traffic in our setup network. This setup is similar to any router on the Internet which is connected with different routers and hosts.

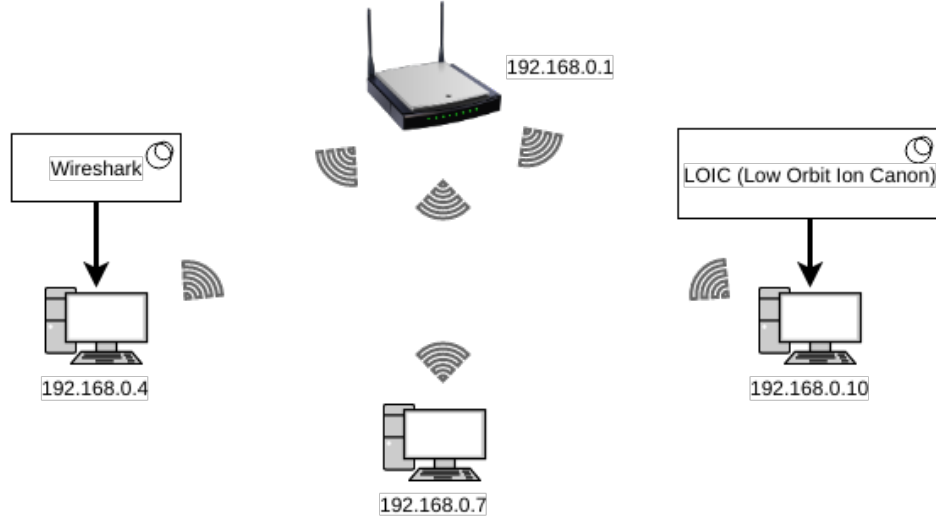


Figure 6: Demo Network: for purpose of demonstration

To collect packets traveling in our network, we start the Wireshark tool. We let it run for a while and then we orchestrate a DDoS attack on one of the host (e.g. 192.168.0.7 in Figure 6) in the network. This attack is engineered using the Low Orbit Ion Canon(LOIC) tool which is installed on one of the host (e.g. 192.168.0.10 in Figure 6) in our network. LOIC is a free tool which is even used by attackers in the real world DDoS attacks. This tool allow us to launch TCP and UDP flood attack on any destination. In this paper we are analyzing the flow based TCP and UDP attack, which is one of the most common type of DDoS attacks.

In the LOIC tool we need to give IP address and the port number of the destination where we want to orchestrate attack while rest of the work is done by it. This tool flood destination with packets and if we choose TCP then it will try to create multiple connections and send packets over them. We start flood attack on one of the system(e.g. 192.168.0.7) in demo network and let it continue for few minuets. We launched such attacks few times in between the packet capturing session.

All the traffic, including the normal and the attack traffic, will get captured using Wireshark. Packets are captured for about five hours in the given network. Once the packets are captured they are saved as pcapng file, which is a Wireshark file format for captured packets. Captured packets during the normal operation and during the attacks are saved separately. The normal packet flow information is used for training and testing the learning algorithms(will be explained in later sections) while attack packets flow information is used for detecting the attack. Wireshark captures every detail of the packets but we don't need all of the information, we are interested only in the IP layer information of the packet. Most of the routers analyze IP layer of the packet

for routing. Having said that there is no reason that other layers of the packet are not analyzed but for demonstration purpose we are analyzing only IP packets.

A data extraction program is written in Python to extract IP layer information from the captured packets. This program extracts address, port and time information from each of the captured IP packets⁸. It also divides the captured data into 300 seconds capture window, thus creating a sample data which is a collection of IP packets captured over the time of 300 seconds. It then write each sample in the separate file for further processing. This sampling of the packets is the continuous process. We then run another program which will extract the flow information form those sample files. one ‘flow’ contain the number of different packets capture for a given destination. We will store this flow information as sample flow. Then we will train learning algorithms using those sample flows. One learning is done those sample files will be discarded and new samples will be created for further training. This training process has to be continuous process in order to correctly reflect the current status of the flow at given router. What ever the new information learned is augmented with the previous learning to have the correct understanding of the flow.

This learning can be done for the time during the day or during the week of year. e.g. We can have separate learning information for flow from morning 9 am to 12 pm and also can have information for evening 6 pm to 12 pm.

Destination IP Address	Protocol	Time stamp(Sec.)	Sample Number
52.6.129.72	6	1512094785.928596000	1
192.168.0.4	6	1512094785.946987000	1
192.168.0.4	17	1512094786.148488000	1

Table 1: Sample file snippet

Flow based model is build as it is more reliable and fast. Packet analyzing is often difficult due to size and encryption. Also destination port number is not a reliable information in detecting attacks because of the fact that attacker use different ports during attack.

Creating a training set for the learning algorithms is an intermediate step in which IP packet count for each destination for a given protocol(e.g TCP, UDP) is calculated. The training set gives us the flow information for each destination (i.e. how many packets are recorded for a given destination IP address during a time window e.g. 300 seconds). We are using Python program to create training set from the sample files. The training set look like following in which each row is one training example with destination IP address as label. A training example is \mathbb{R}^3 vector whose elements are the number of packets observed for a particular protocol during a 300

⁸packets containing IP information

second time.

Destination IP Address	IP Packet count		
	ICMP	TCP	UDP
172.217.10.134	0	8	12
65.19.96.252	5	0	192
68.67.178.134	0	78	0

Table 2: Training Set with three training examples

We ran Python program to convert each sample file to training set and testing sets. Each sample file has corresponding train/test set file. We store both train and test sets on the file systems so that they can be used for training and testing the algorithms.

There are around 150 protocols managed and assigned by the Internet Assigned Numbers Authority (IANA) but most commonly used protocols in the DDoS attack are ICMP, TCP and UDP protocols. For the training and analysis purpose we are using only these three protocols as the desired feature. In the larger system such as routers managed by ISP, other protocols can also be used as features if required.

5.1 Machine Learning

According to Tom Mitchell (1998) A computer program is said to learn from experience E with respect to some task T and some performance measure P , if its performance on T , as measured by P , improves with experience.

A learning algorithm build the hypothesis using training set as input then that hypothesis is used to perform predictions. Most common categorization of machine learning algorithms is Supervised and Unsupervised.

Let f be the function which we need to guess from input vector $X = \{x^1, x^2, ..., x^m\}$. This X also called as training set where x 's represent an 'input variable' or a 'feature vector' and m is number of feature vectors in the training set. Let h be the hypothesis about the function f . $h \in H$ and $f \in H$, where H is class of some functions. Both f and h can be vectors. We select h based on a training set, Ξ , of m input vector examples. In Supervised leaning we know the values of f for m samples in the training set Ξ . We assume that if we can find a hypothesis, h , that closely agrees with f for the members of Ξ , then this hypothesis will be a good guess for f when Ξ is large. In Unsupervised learning, we simply have a training set of vectors without function values for them. The problem in this case, typically, is to partition the training set into subsets, $\Xi_1, ..., \Xi_R$, in some appropriate way. [12]

Supervised algorithm such as One Class Support Vector Machine(One Class SVM) [13] could

be efficient to identify the anomalies in the data but this algorithm is process and memory intensive, hence training the algorithm for each and every IP address is very costly in terms of resources. Because of the resource constraints of the router, our approach is to first cluster the IP addresses based on the features using Unsupervised learning algorithms such as k-means and then apply One Class SVM on the clusters to decide on the boundaries of those clusters. k-means algorithm is fast and consume less resource compared to One Class SVM, which make them good to be used on the devices such as router which has less processing power and less memory.

5.1.1 Feature Scaling

Before feeding data to learning algorithm, we have to do feature scaling, also called Standardizing. Feature scaling is done by removing the mean and scaling the feature to a unit variance value. It is necessary because of the fact that, different features which are at the different scales could cause one feature dominating the other in the algorithm output result. e.g. consider two vectors (1, 2, 3000) (1, 3, 2000). If we calculate the Euclidean distance between these two vectors then it will be square root of $(1 - 1)^2 + (2 - 3)^2 + (3000 - 2000)^2$. From this, it is evident that the larger term is dominating the result.

First we will convert training examples into a vector. Each row of the vector is one training example and each element in the vector is the feature. To standardize the input vector, mean and standard deviation is calculated for each feature in the input vector. Then new vector is created by subtracting the mean from every element of the feature vector and then dividing values of each feature vector by its standard deviation. The new vector created after this step is standardized vector, which is used as input to the learning algorithms.

$$\text{Standardization formula: } x' = \frac{x - \bar{x}}{\sigma}$$

Where x is the feature vector, \bar{x} is the mean and σ is its standard deviation.

5.1.2 Clustering

k-means clustering is one of the most efficient algorithm for creating clusters. The k-means problem is to find set of k points $\{\mu_1, \mu_2, \dots, \mu_k\}$ called centroids, $\mu_k \in \mathbb{R}^d$ and $K \in \mathbb{N}$, for training set $X = \{x^1, x^2, \dots, x^m\}$, where $x^i \in \mathbb{R}^d$ and $i = 1, 2, \dots, m$ is a training example, such that the mean square distance from each training example x^i of training set, to its nearest center μ_k ($\text{argmin}_k \|x^i - \mu_k\|^2$), is minimum. $m \in \mathbb{N}$, are the number of training examples. [14]

Following is Lloyd's algorithm which is most popular heuristic algorithm for k-means clustering. The clustering that we will be doing is of destination IP addresses, such that each cluster will have some number of destination IP addresses.

Algorithm 1 Lloyd's k-means algorithm

Arbitrarily initialize K cluster centroids $\mu_1, \mu_2, \dots, \mu_k$

repeat

for $i \leftarrow 1, m$ **do**

$$c^i = \operatorname{argmin}_k \|x^i - \mu_k\|^2$$

▷ Cluster Assignment

end for

for $k \leftarrow 1, K$ **do**

$$\mu_k = \frac{1}{n} [x^{(k_1)} + x^{(k_2)} + \dots + x^{(k_n)}]$$

▷ Move Centroid

end for

until convergence

For this paper we will be using the k-means++ algorithm [15] which is an improvisation of k-means, where arbitrary initialization step is replaced by following simple, randomized seeding technique. This k-means++ algorithm is $O(\log k)$ -competitive with the optimal clustering.

Let $D(x)$ denote the shortest distance from a data point to the closest center we have already chosen, then following is the k-means++ algorithm.

Algorithm 2 k-means++

1: Take one center μ_1 , chosen uniformly at random from X .

2: Take a new center μ_k , choosing $x \in X$ with probability $\frac{D(x)^2}{\sum_{x \in X} D(x)^2}$

3: Repeat Step 2. until we have taken k centers altogether

4: Proceed with the Lloyd's k-means algorithm skipping Arbitrarily initialize stage

For our DDoS attack detection program, we will be using the Scikit-learn libraries. Scikit-learn is the most popular and rich open source machine learning software library for the Python programming language and they have implementation for both k-means and One Class SVM machine learning algorithms, also they have data preprocessing programs such as feature scaling.

We will be using training set in the format given in the figure 2 and we will be using k-means++ Scikit-learn library for clustering.

Before doing the clustering, we have to first determine the number clusters and centroids⁹. Deciding on the number of clusters is important, because randomly choosing the number of cluster will not be useful to have correct clustering. We will use the Elbow method to find the optimal number of clusters. The Elbow method check the percent of variance explained as function of the number of clusters. Following are the Elbow Diagrams for four samples.

⁹Centroid is the vectors which is arithmetic mean of all the point in the cluster.

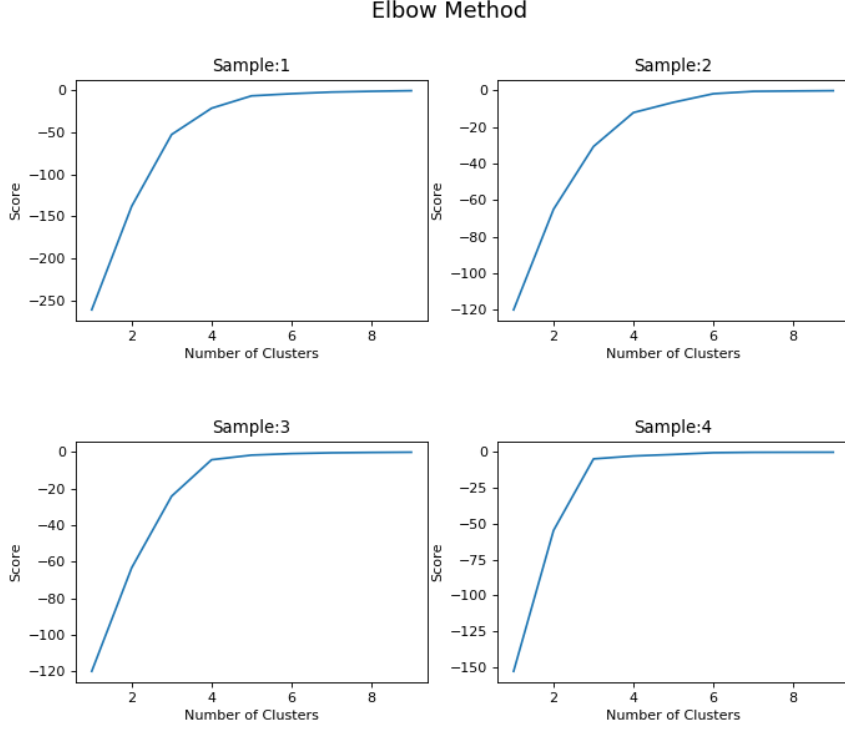


Figure 7: Elbow Method for cluster count detection

Using the elbow method, variance for each cluster number is calculated and the cluster number which produces less variance for the next cluster number is selected as best choice for the give sample data.

To have correct clustering we first run the k-means algorithm to determine the central vectors called centroids. Clusters are represented by these centroids. As we have multiple ‘samples’ (sample is the flow information) so we have cluster centroids for each sample. We will save all the centroids and then find the median of all the centroids removing the outliers so that we have good estimate of the centroids. This estimated centroids are used for clustering the samples in the future. Following is the example of centroid vector where rows represents cluster label and columns represent the feature.

Cluster	ICMP	TCP	UDP
0	-0.16815612	-0.14928111	-0.16948046
1	-0.18181818	5.13527652	5.68956244
2	5.08663322	-0.27110845	-0.099885
3	-0.18670401	-0.18804342	-0.018538

Table 3: k-means cluster centroids

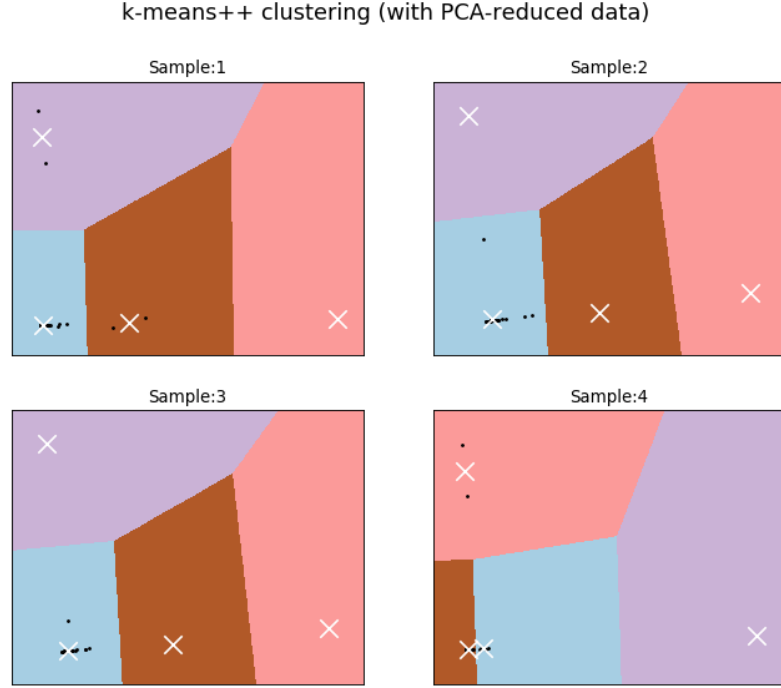


Figure 8: Clustering using k-means++ algorithm

To draw data points on 2 dimension space we had to reduce the 3 dimensional training example into 2 dimension without losing much information. We have achieved this by using Principal-Component Analysis (PCA). PCA is a technique for taking a dataset consisting of a set of tuples representing points in the high-dimensional space and finding the directions along which the tuples line up best. [16]. We have used Scikit-learn PCA module for this purpose.

Using the centroids obtained in the previous step we ran k-means++ algorithm and then test it for accuracy. k-means++ algorithm gives label to each training example. This label is the cluster number to which that training example has been assigned to. So the new labeled data look like following.

Destination IP Address	ICMP	TCP	UDP	Cluster
172.217.10.134	0	8	12	1
65.19.96.252	5	0	192	0
68.67.178.134	0	78	0	2

Table 4: Labeled Training Set (with cluster number)

We ran the k-means++ for both training and testing set and produce labels for both. We then checked how similar the test set clustering is with train set clustering i.e. we are checking

if the IP address has the same cluster number assigned in both training and test data. For measuring this similarity Rand Index(RI) [17] is used. RI is a measure of how many percent does the test clustering matches with the trained model.

$$RI = \frac{TP+TN}{TP+FP+FN+TN}$$

where TP, TN, FN and FP are the number of true positives, true negatives, false positives and false negatives respectively.

We have observed that with more number of training sets RI index improves.

Because of the reason that the training sets contain the flow information for different IP on a router during window of 300 seconds, there is a very high possibility that the same destination IP address is captured in multiple training set. Our goal is to find out the correct cluster for the destination IP address, and to achieve this goal, we check all the cluster assignment for a given destination IP address and we select the cluster which has the highest occurring frequency among all the training sets. We also count the average number of packets going to a given IP address which will help us reduce the error in detecting anomaly. We will save this labeling information and cluster count information for each IP destination on the file system.

Destination IP Address	Cluster	Packet Count
74.125.141.106	1	113
72.30.2.182	0	16
64.94.191.14	0	22

Table 5: Learned information after clustering

This information tell us the normal behavior of the packets traveling from the router to a given destination. As we have fixed the centroids for the clustering algorithm, every time in the future we should expect the IP address to be found in the same cluster if flow of packets for a given destination is as per the previous knowledge. If there is a DDoS attack on a any destination with flooding, then we can expect to see the destination IP address assigned to different cluster.

But from the experiments on different data sets it is found that the destination under attack is labeled with the same cluster number when it was not under attack. This happened because there is no other cluster it can assigned to, so it gets assigned to cluster whose centroid is nearest.

To avoid such situation, we will have to create the boundaries for the clusters. This scheme will make sure that the attack on the destination IP address will be detected even if the destination is assigned to the same cluster it was labeled in the past.

To create cluster boundaries we have used One Class SVM, a supervised machine learning algorithm. Compute and storage requirements of SVM increase rapidly with the number of training vectors, because SVM is a quadratic programming problem (QP). So the approach

presented in the paper is more efficient because the number of cluster are limited and always be far less than the number of destination IP address. Training one class SVM on clusters will be far less process and memory intensive than training on massive number of IP address.

For example the number of clusters we have create in our analysis are 4 in number, while the number of IP addresses to which packets are flowing from our router are 268, which is more than 60 times the count of clusters we have created.

5.1.3 Anomaly Detection using One Class Support-Vector Machine

Support-Vector Machine: Support-Vector Machine is a supervised learning algorithm which tries to classify data. Classification is based on the label of the training set. Consider the training set $\{(x^1, y^1), (x^2, y^2), \dots, (x^m, y^m)\}$, where $x^i \in \mathbb{R}^d$ is the training example and $y^i \in \{-1, +1\}$ is the label which is the classification value for x^i ,

SVM try to create non-linear separation boundary by projecting data points through a non-linear function ϕ to higher dimension space. The data points in space I which can not be separated by a line are projected to the feature space F where there can a hyperplane that separate data point of one class form another. If the hyperplane is projected back on original space I then we get non-linear curve. [13]

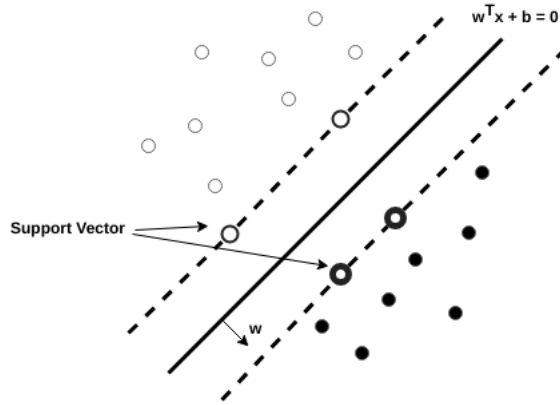


Figure 9: Linear Separation using Support Vector

The hyperplane is represented with the equation $w^T x + b = 0$, where $w \in F$ and $b \in R$. This hyperplane separates the training example labeled with -1 and 1 into different classes. The position of the hyperplane is such that the distance from the closest point from each class to the hyperplane is same. To avoid the over-fitting, slack variables ξ^i are introduced. Over-fitting happen because the learned hypothesis fit training examples so well that it fails to generalize the new examples. The constant $C > 0$ is the regularization parameter. If C is chosen large, miss-classification of training examples can be avoided. If chosen small, then we may miss-classify

few examples, but the margin will be large, so that most of the points will be far away from the decision boundary. The SVM optimization problem is stated as follows. [13] [18]

$$\begin{aligned} \min_{w,b,\xi^i} \quad & \frac{\|w\|^2}{2} + C \sum_{i=1}^m \xi^i \\ \text{subject to:} \quad & \\ & y^i(w^T \phi(x^i) + b) \geq 1 - \xi^i \text{ for all } i = 1, \dots, m \\ & \xi^i \geq 0 \text{ for all } i = 1, \dots, m \end{aligned}$$

If this minimization problem is solved using Lagrange multipliers then the classification function $f(x)$ can be stated as.

$$f(x) = \text{sgn}(\sum_{i=1}^m \alpha^i y^i K(x, x^i) + b)$$

α^i here are the Lagrange multipliers and x^i with α^i are called the Support Vectors.

The function $K(x, x^i) = \phi(x)^T \phi(x^i)$ is known as the kernel function. kernel function responsible for projecting data points to the hyperplane.

One Class Support-Vector Machine: One Class Support-Vector Machine (One Class SVM) is the extension of SVM which detect boundaries of the training set so that every new training example will be classified as belong to training set or not. It separates all the training set data point from feature space F and maximizes the distance of hyperplane from F . This creates a binary function which returns +1 for the training example that fits in the trained set region, otherwise it will return -1.

The minimization function of One Class SVM is slightly different than the SVM. [13]

$$\begin{aligned} \min_{w,\rho,\xi^i} \quad & \frac{\|w\|^2}{2} + \frac{1}{\nu n} \sum_{i=1}^m \xi^i - \rho \\ \text{subject to:} \quad & \\ & (w \cdot \phi(x^i)) \geq \rho - \xi^i \text{ for all } i = 1, \dots, m \\ & \xi^i \geq 0 \text{ for all } i = 1, \dots, m \end{aligned}$$

The new variable introduced ν in place of C in previous SVM equation is used to set upper bound on outliers/anomalies and lower bound on the number of training examples.

We are using Scikit-learn's 'OneClassSVM' library to train the model and create classifier for each cluster. The input vector to the classifier is the set of all the training examples belonging to same clusters. Thus, if we have four clusters then we will have four classifier. The input vector to the classifier will be of the form shown in Table 4.

Following is the result of modeling on the training data sets. Each cluster has its own model.

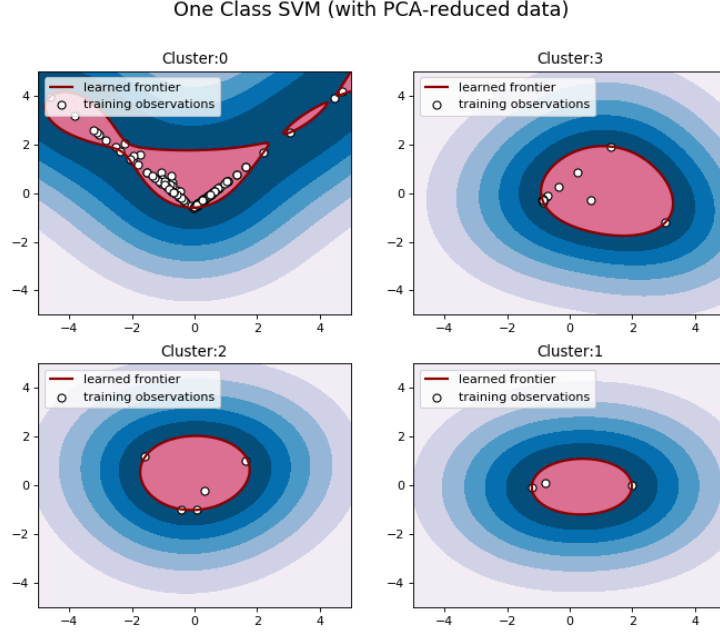


Figure 10: One Class SVM

6 Detection

By this stage, we have all the destination IP address, at the router, labeled with cluster and their count of average number of packets during the flow. Also we have classifier models for each of the cluster.

To detect the attack we capture the flow in the specified interval (e.g. 300 seconds, in which we have trained our models). We transformed the captured flow into a test set. The created test set is then clustered using already centroids which were calculated for the clustering of the training sets. Now we have each IP in the new test set labeled with the cluster number. If there is any destination IP address for which new cluster label does not match with the already known cluster label then that destination address added to the suspect list. If new label matches with the old then the feature vector for that destination address is passed to the cluster classifier to check its subscription the cluster its found in. i.e. if an IP address is labeled with cluster number 0 then we use One Class SVM classifier for cluster number 0. If the output of the classifier is -1 then then that destination is added to the suspect list. For every destination from this compiled list, the number of packets observed is also recorded. If there is significant difference between the number of capture packet in the past and in the present then that destination is recorded as DDoS attack candidate because failing to found in the same cluster boundary it was in the past is a sign of the change in th behavior of the traffic for destination IP address on the router

where this analysis is done.

Using this we have successfully detected attack on the destination IP address 92.168.0.7 in our modeled network.

Next task of the router then will be to communicate the change in the behavior of traffic observed at that router for a that destination. The destination address here could be a router or an application server. Destination system i.e. either router of the server will collect the information received. This information can be used by destination to know the nature of the attack.

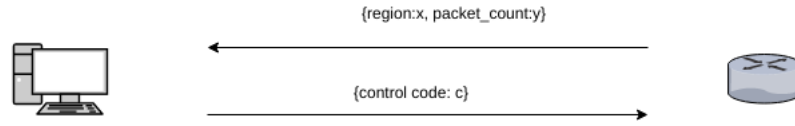


Figure 11: Router Network Communication

To Communicate the destination network about the change in behavior, router can use the existing ICMP protocol. ICMP protocol is used to provide feedback about problems in the communication environment. ICMP messages are sent in several situations: for example 1) when a datagram cannot reach its destination. 2) when the gateway does not have the buffering capacity to forward a datagram. 3) when the gateway can direct the host to send traffic on a shorter route. [10] Similarly we can use ICMP protocol to inform destination system about the change in the traffic. ICMP protocol has many unused type code (there can be 0-255 types but as of now only 0-41 are in use) available. We can create a new ICMP ‘type’ to send DDoS detection information from the router to the destination system and then destination system can send mitigation instructions to the routers.

Depending on the type and the severity of the situation, the destination system can decide whether or not to inform router to block the traffic coming from that router. We will let this decision to be taken at the destination system. There can be different parameter based on which the destination system can decide blocking the flow. Many questions can be asked before making the decision, such as how many routers have reported the change in traffic?, Is the attack information coming from the region which never had traffic in the past?

7 Conclusion

A novel way to identify and mitigate DDoS attack is discussed in this paper. With the advance of NVF(Network Virtual Functional) it is easy to push the learning algorithms to the routers allowing them to detect a DDoS attacks using machine learning algorithms. The DDoS detection and mitigation information can be communicated using existing ICMP protocol making system available to every one who want to use it. This approach has been tested with the small network so further experiments are needed to be performed on the larges networks.

References

- [1] Derek Kortepeter. Destructive ddos attacks increasing at a rapid rate. <http://techgenix.com/ddos-attacks-increasing/>, December 2017. Online; accessed 22-August-2017.
- [2] D. W. Davies. A communication network for real-time computer systems. *Radio and Electronic Engineer*, 37(1):47–51, January 1969.
- [3] Cisco Security portal. A cisco guide to defending against distributed denial of service attacks. <https://www.cisco.com/c/en/us/about/security-center/guide-ddos-defense.html>. Online; accessed 1-August-2017.
- [4] S. M. Bellovin. A look back at "security problems in the tcp/ip protocol suite". In *20th Annual Computer Security Applications Conference*, pages 229–249, Dec 2004.
- [5] Ken Dunham and Jim Melnick. *Malicious Bots: An Inside Look into the Cyber-Criminal Underground of the Internet*, chapter 1. Introduction to Bots. CRC Press, August 2008,.
- [6] Network Working Group. Traffic flow measurement: Architecture. <https://www.ietf.org/rfc/rfc2722.txt>, October 1999.
- [7] Cisco. Introduction to cisco ios netflow. https://www.cisco.com/c/en/us/products/collateral/ios-nx-os-software/ios-netflow/prod_white_paper0900aecd80406232.html, May 2012. Online; accessed 22-August-2017.
- [8] sFlow. Using sflow. http://www.sflow.org/using_sflow/index.php. Online; accessed 16-November-2017.
- [9] Cisco. What is a network switch vs. a router? <https://www.cisco.com/c/en/us/solutions/small-business/resource-center/connect-employees-offices/network-switch-what.html>. Online; accessed 16-November-2017.
- [10] Network Working Group. Icmp. <https://tools.ietf.org/html/rfc792>, September 1981.
- [11] Omar Santos. The size of the internet global routing table and its potential side effects. <https://supportforums.cisco.com/t5/network-infrastructure-documents/the-size-of-the-internet-global-routing-table-and-its-potential/ta-p/3136453>, May 2014. Online; accessed 22-August-2017.
- [12] Jeffrey D. Ullman Jure Leskovec, Anand Rajaraman. *Mining of Massive Datasets*, chapter Clustering. Stanford University, 2014.

- [13] Bernhard Schölkopf, Robert Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, pages 582–588, Cambridge, MA, USA, 1999. MIT Press.
- [14] Nathan S. Netanyahu Christine D. Piatko Ruth Silverman Tapas Kanungo, David M. Mount and Angela Y. Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE*, 24(7):881–892, July 2002.
- [15] David Arthur and Sergei Vassilvitskii. K-means++: the advantages of careful seeding. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2007.
- [16] Jeffrey D. Ullman Jure Leskovec, Anand Rajaraman. *Mining of Massive Datasets*, chapter 11. Dimensionality Reduction. Stanford University, 2014.
- [17] William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.
- [18] Jeffrey D. Ullman Jure Leskovec, Anand Rajaraman. *Mining of Massive Datasets*, chapter 12. Large-Scale Machine Learning. Stanford University, 2014.