On the Characteristics of Persistent Communities of Enterprises

Han Zhang, Mark Teodoro, Christos Papadopoulos, Allison Mankin

Colorado State University, Verisign Labs
zhang@cs.colostate.edu,mteodoro@verisign.com,christos@cs.colostate.edu,
amankin@verisign.com

Abstract. The set of hosts that communicate with an enterprise is a complex mix. Some may visit the enterprise occasionally and never come back, others communicate with the enterprise very frequently. We define the latter as *Persistent Hosts* and the set as a *Persistent Community*. Characterizing *Persistent Communities* benefits enterprises in several ways, including security, network management, traffic engineering and more. In this paper, we use 78 billion flow records collected from a sample of 84 enterprises for an entire month to explore the characteristics of *Persistent Communities*. First we characterize the *Persistent Community* of each enterprise and find that for 90% of the enterprises, less than 21% of the hosts are persistent, yet they contribute more than 50% of the traffic. Then we correlate the *Persistent Communities* between multiple enterprises. We find that as the number of enterprises that a *Persistent Host* communicates with increases, their communication hours each day also increases. Moreover, we correlate *Persistent Communities* with DDoS attacks and find that while some *Persistent Hosts* are involved in UDP-based DDoS attacks, they only contribute a small portion of the overall attack traffic. Based on our findings, we give a simple case study of using *Persistent Communities* to detect business changes of enterprises, prioritize traffic and perhaps result in improved DDoS protection.

Keywords: Persistent Community, Enterprise Network, DDoS Attack

1 Introduction

For each enterprise, there are a lot of outside hosts that communicate with it every day. And the hosts inside the enterprise also visit many outside websites, data centers, and so forth. This set of outside hosts is a complex mix and it changes every day. Some outside hosts may visit the enterprise's networks once and not return. On the other hand, there are some hosts that communicate with the enterprise almost every day. In the context of this study, we define a *Persistent Host* as the outside host that communicates with an enterprise network frequently (e.g., seven days a week), and define a *Persistent Community* as a set of these *Persistent Hosts*. In this paper, *Persistent Community* and *Persistent Hosts* can be used interchangeably.

The *Persistent Hosts* can be categorized into two groups. The first group includes the hosts that have persistent interests in the enterprise, like the people who read news from *The New York Times* website every day. Besides, for the enterprises that provide services (e.g., cloud storage), their customers might communicate with them very often. The second group includes the hosts that the enterprises are interested in. This group could be composed of the servers providing popular or critical services (e.g., AV patches, cloud storage), the websites visited by enterprises' employees frequently (e.g., Facebook, Google), and so forth. We do not attempt to distinguish these two groups because we believe they are both important to the enterprises. Group 1 benefits enterprises' business while group 2 is necessary or critical to enterprises' daily operations and employees' work.

In this paper, we study the *Persistent Hosts'* characteristics and also investigate whether they participate in DDoS attacks by using real attack data. Measuring *Persistent Communities* is not only interesting on its own, but it also benefits the enterprises in several ways, including network management, traffic engineering, security and more. For example, we can use *Persistent Hosts* as whitelist to improve DDoS protection systems. The profile of a *Persistent Community* can also be used for capacity allocation. Moreover, higher priority can be assigned to the *Persistent Community* than the others during periods of heavy traffic.

Our study is limited because some of the flow records in our data set are collected after packet sampling, which introduces two issues. One is that some short flows may be missed, thus we don't consider all the hosts that communicate with the enterprises, and the other is that the traffic volume has to be estimated. Even with these limitations, there is much to explore in our data set, which covers an entire month including 78 billion flow records between 7.4 million hosts inside a sample of 84 enterprises and 500 million external hosts. Notice that the traffic seen in a specific enterprise may differ from others. However, the enterprises included in our data set provide great variety, including banking, web hosting, retail, telecommunications, and so forth. Consequently, we expect they are representative for many enterprises. To the best of our knowledge, this is the first study to explore *Persistent Hosts* for such a large number of real enterprise networks. The contributions of our paper are three-fold:

1. Characterize the *Persistent Community* for single enterprise from several aspects, including its size, the volume of traffic it contributes, and its country distribution.
2. Correlate the *Persistent Communities* between multiple enterprises and characterize them from various perspectives, including the organizations that they belong to, communication frequencies and the relationship between the size of a *Persistent Community* and the number of enterprises that it persistently communicates with.
3. Correlate real DDoS attack data with *Persistent Hosts* of the victim enterprises and investigate whether they perform DDoS attacks. We find that

while some of the *Persistent Hosts* are involved in UDP-based DDoS attacks, they only contribute a small portion of attack traffic.

The reminder of this paper is structured as follows. Section 2 describes the data sets used in our experiments. In Section 3, we characterize the *Persistent Community* for a single enterprise from various aspects. Then we correlate the *Persistent Communities* for multiple enterprises and show their characteristics in Section 4. Section 5 investigates whether the *Persistent Communities* perform real attacks. We offer some suggestions on using *Persistent Communities* to improve DDoS protection systems, prioritize traffic and more in Section 6. Related work is presented in Section 7. Finally, we introduce future work in Section 8 and draw conclusions in Section 9.

2 Dataset

2.1 Flow Records

NetFlow [5], [6] was introduced on Cisco routers to collect flow information of IP network traffic. Here a flow is defined as a set of packets that share common attributes, for example, a unidirectional sequence of packets that have the same ingress interface, source IP address, destination IP address, IP protocol, source port, destination port and type of service. As network traffic enters or exits a router interface, the router can generate the NetFlow records for the traffic and then export the records to a specified collector. Internet Protocol Flow Information Export (IPFIX) [7] is an IETF protocol designed based on NetFlow version 9, and it defines how IP flow information is to be formatted and transferred from an exporter to a collector. sFlow [3], [8] is a sampling technology used to capture traffic statistics from the device it is monitoring, like switch and router. The common thing shared by these three techniques is that they all tell who talks to whom, and how much data is exchanged between them (might be sampled). Many companies collect NetFlow, IPFIX and sFlow data for their customers and use it to improve the provided services. For example, ISPs use the collected data for network traffic accounting, usage-based network billing, network planning, and more [5]. The security service providers can collect the flow data from the customers' edge routers to monitor the traffic and provide DDoS protection services.

Verisign provides cloud-based DDoS protection services to many enterprises around the world. Specifically, Verisign collects sample packets, NetFlow records and other pertinent information from switches, routers and other devices, and feed this information into a detection engine for threat detection, alerts and reporting [1]. This data set includes the NetFlow, IPFIX and sFlow data for the incoming traffic to a sample of 84 enterprise networks from March 1st to March 31st, 2014. The edge routers of the enterprises generate NetFlow/IPFIX/sFlow records and export them to the collector, then the data is stored on machines.

We call this data set *OneMonthFlows*. It includes 78 billion flow records between 7.4 million hosts inside enterprises and 500 million external hosts. Due to the high volume of traffic and limited computation resources, some of the routers that generate flow records have to perform packet sampling. Some statistics of sample ratios and protocols for this data set are listed in Table 1a and Table 1b respectively. In some experiments of this paper, we only investigate the hosts that communicate with enterprises within one week, so we extract the flow records from March 11th to 17th from *OneMonthFlows*, store them in a separate data set and name it as *OneWeekFlows*.

Sample Ratio	Flows	Percentage
1	6.24E+10	79.8%
1000	9.21E+9	11.8%
512	6.18E+9	7.9%
2048	2.50E+8	0.3%
Others	1.78E+8	0.2%
Total	7.82E+10	100%

(a) Sample Ratio Statistics

Protocol	Flows	Percentage
TCP	6.48E+10	82.8%
UDP	1.14E+10	14.5%
ICMP	1.00E+9	1.3%
ESP	9.02E+8	1.2%
Others	1.80E+8	0.2%
Total	7.82E+10	100%

(b) Protocol Statistics

Table 1: Data Set *OneMonthFlows* Statistics

2.2 DDoS Attack Alerts

As mentioned in Section 2.1, Verisign provides DDoS protection services to many enterprises around the world [1]. As many security service providers do, Verisign also analyzes these enterprises' network traffic and logs the triggered DDoS alerts. This data set includes the DDoS alerts for these enterprises from March 1st, 2014 to March 31st, 2014, named as *DDoSAlerts*. Each alert contains some basic information about an attack, like the start time, end time, type of attack and the victim enterprise id (not the enterprise name). We use *DDoSAlerts* as the ground truth to investigate whether *Persistent Hosts* perform attacks, which will be covered in Section 5.

One thing we want to emphasize is that the data sets *OneMonthFlows*, *OneWeekFlows* and *DDoSAlerts* are only available to certain authorized Verisign employees. The main author was authorized to use these data while working on an internship to improve the services that Verisign provides to customers.

3 Single-Enterprise Persistent Communities

3.1 Overview

In this section, we give an overview of the hosts that communicate with single enterprise every day. In particular, we use the data set *OneMonthFlows* to measure the number of hosts that communicate with each enterprise for multiple sequential days starting from March 1st. We denote the number of hosts that communicate with enterprise i for j sequential days as S_{ij}. As S_{ij} varies greatly depending on enterprises' scales and business types, we normalize it by dividing the number of hosts appearing for x sequential days by the number of hosts appearing for 2 sequential days, calculated as S_{ij}/S_{i2}. The results are shown in Figure 1. The x-axis is the number of days that the hosts appear sequentially (j), y-axis is S_{ij}/S_{i2} and different lines stand for various enterprises. Here we only consider the enterprises that communicate with more than 500 different hosts on each day. From the figure we observe two things:

1. For almost all the lines, they drop very slowly after seven days. In other words, the hosts that communicate with an enterprise for seven sequential days are likely to continue communicating beyond that period.
2. There is one clear outlier in the figure, which is the second line on the top that drops significantly at day 13. The reason is that there are 41915 hosts that communicate with this enterprise for 12 sequential days but only 278 for 13 sequential days. In fact 5 edge routers for this enterprise exported flow records to us from March 1st to March 12th, but only 4 after March 13th.

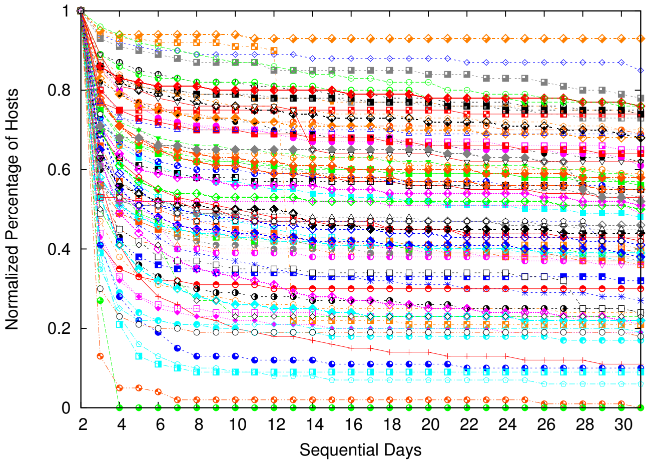


Fig. 1: Hosts Communicating with Enterprises for Multiple Sequential Days

3.2 Definition of a Single-Enterprise Persistent Community

In the last section, we found that the hosts communicating with an enterprise for seven sequential days will continue communicating beyond that period. Consequently, we decided to use seven days to define a *Persistent Host*. Another reason to use seven days is because it covers an entire week, both workdays and weekends. We define the host that communicates with a certain enterprise at least N days out of seven days as a *Single-Enterprise Persistent Host* and define a *Single-Enterprise Persistent Community* for this enterprise as the set of these hosts. Notice that the N days don't have to be sequential.

3.3 Size of a Single-Enterprise Persistent Community

In this section and Section 4, we focus on the data within one week and use the data set *OneWeekFlows* for analysis. First we extract the *Single-Enterprise Persistent Community* for each enterprise by using N from 4 to 7. Due to different enterprises' scales, the sizes of their *Persistent Communities* may vary greatly. Consequently, we consider the ratio of the size of a *Persistent Community* to the total number of hosts that communicate with the same enterprise. We use n_{ij} to denote the size of a *Persistent Community* for enterprise j at day i, and use N_{ij} to denote the number of all the hosts that communicate with enterprise j at day i. Then we calculate *percentage of persistent hosts* for enterprise j as:

$$p_j = \sum_{i=0}^{6} n_{ij} / \sum_{i=0}^{6} N_{ij}$$

After calculating p_j for all the enterprises, we plot their cumulative distribution function (CDF), as shown in Figure 2. The x-axis is the *percentage of persistent hosts* and y-axis is the cumulative possibility. Different lines stand for various N. If we pick the point (10, 0.65) on the left most line, it means that by using 7 as N, there are 65% of the enterprises whose *percentage of persistent hosts* is less than 10%. From the figure we learn two things:

1. For most of the enterprises, a small portion of the hosts that communicate with them are persistent. For example, if we use 7 as N, 90% enterprises' *percentage of persistent hosts* is less than 21%.
2. As N decreases, the line moves right. The reason is that by using a small N, more hosts are considered as persistent, thus the *percentage of persistent hosts* gets greater.

Besides *Single-Enterprise Persistent Communities*, we also define *Single-Enterprise Persistent Blocks* as the blocks of hosts that communicate with a certain enterprise at least N days out of seven days. Currently we consider /24 CIDR blocks. We calculate the *percentage of persistent blocks* similarly with *percentage of persistent hosts* and show the CDF in Figure 3. The figure tells that in general the *percentage of persistent blocks* is greater than *percentage of persistent hosts.* The reason is that the hosts belonging to the same /24 subnets are aggregated, so they are more likely to appear for multiple days and considered as persistent.

3.4 Traffic Volume from a Single-Enterprise Persistent Community

In the last section, we have quantified the *Single-Enterprise Persistent Community* using various values as N. In the remaining part of this paper, we use 7 as N,

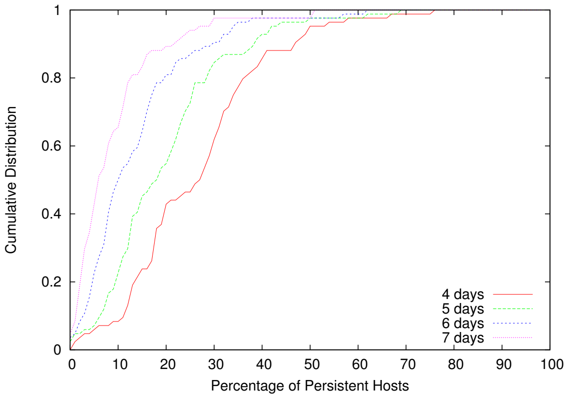


Fig. 2: CDF of Persistent Community

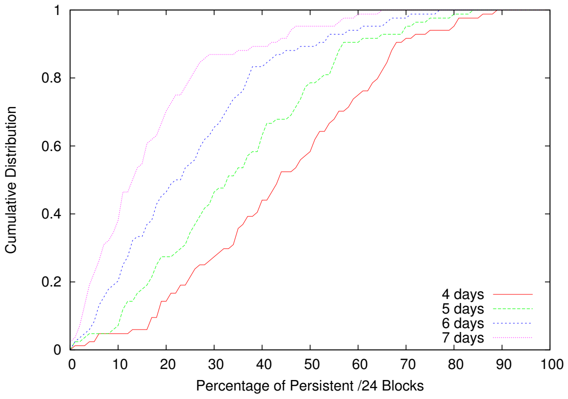


Fig. 3: CDF of Persistent Blocks

which means the *Persistent Community* has to appear every day for a week. The results in the last section show that the *Persistent Community* is a small part of all the hosts that communicate with the enterprises. As we know, a small number of hosts could contribute a great amount of traffic. In this part we investigate how much traffic the *Persistent Communities* contribute. Notice that the aim of this section is not to calculate a specific number of traffic volume contributed by *Persistent Hosts*. Instead, we want to provide a general sense that whether *Persistent Hosts* contribute a lot or a small portion to all the traffic being sent to the enterprises. Specifically, we divide the number of packets sent from the *Persistent Community* to the enterprise by the number of all the packets sent to the enterprise, named as *percentage of persistent community packets*.

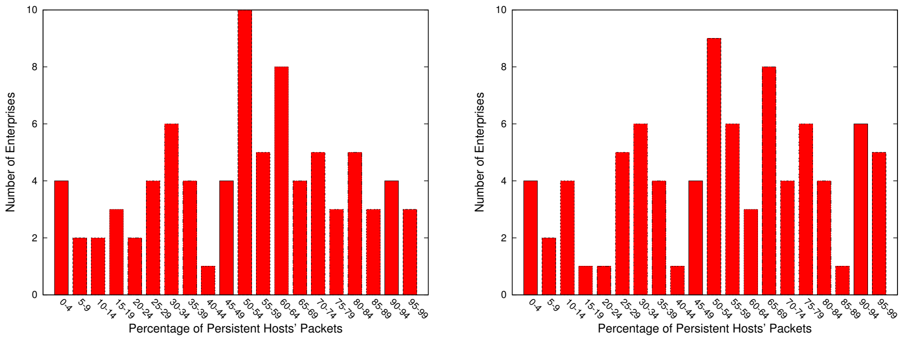
One challenge here is that some of the flow records in *OneWeekFlows* are collected after packet sampling. We deal with this issue by two means. Recall that 80% of the flow records are collected without packet sampling, so the first method is to only use these unsampled data for analysis. The results are shown in Figure 4a, x-axis is the *percentage of persistent community packets*, and y-axis is the number of enterprises. For example, the left most bar means that there are 4 enterprises whose *percentage of persistent community packets* is between 0% and 5%. From the figure we find that for 50 out of 82 enterprises, their *percentage of persistent community packets* is greater than 50%. If we consider the traffic for all the enterprises, 59.4% of the total traffic is sent from the *Persistent Communities*. The reason why we miss two enterprises is that all of their flow data is packet-sampled.

The second method is to use all the flow data including sampled and unsampled for analysis. In this case, we need to estimate the amount of traffic for the sampled data. We decided to make a rough estimate by simply multiplying the number of packets by the sample ratio. There are some works studying the accuracy of such an estimation. Barakat [21] finds that the packet sampling misranks the largest flows and introduces a new protocol-aware ranking method to improve the accuracy. In [22], the authors study the impact of packet sampling on flow attributes. They find that packet sampling has great impact on data rate while very small on packet rate. These works focus on individual flow and use a short period of time as time window, usually less than 5 minutes. We agree with these works that packet sampling has impact on estimation of flows' packets. However, instead of individual flow, we focus on communications between a host outside an enterprise and all the hosts within this enterprise's network during a very long period of time - one day, which could be composed of thousands of flows. More than that, we estimate the number of packets between a certain enterprise and its entire *Persistent Community*, which is a large group of *Persistent Hosts*. Consequently, we are estimating a great amount of flows during a long period of time. Moreover, we study the relative value instead of the exact number of packets. Specifically, in the following experiments we consider the ratio of the number of packets sent by *Persistent Community* to the total number of packets sent by all the hosts to a certain enterprise, but not the exact number of packets. Based on these reasons, we expect that packet sampling should have a smaller impact on our work than on individual flows.

The results of considering all the flow records are given in Figure 4b. This figure is a little different from Figure 4a, but still tells us that for 52 out of 84 enterprises the *percentage of persistent community packets* is greater than 50%. If we consider the traffic for all the enterprises, 70.9% of the total traffic is sent from the *Persistent Communities*. In summary, the results of both methods indicate that *Persistent Hosts* contribute a lot to all the traffic sent to the enterprises.

3.5 Country Distribution of Single-Enterprise Persistent Communities

In this section, we describe the country distribution of *Single-Enterprise Persistent Hosts* and investigate whether they match the countries that have the most hosts communicating with the enterprises. For each *Persistent Host*, we use *MaxMind* [2] to check the country it belongs to. First we merge the *Persistent*



(a) Flow Records No Packet Sampling

(b) All Flow Records

Fig. 4: Contribution of Persistent Hosts' Packets

Hosts for all the enterprises and list the countries having most *Persistent Hosts* in Table 2. However, it could be the case that some enterprises have a great number of *Persistent Hosts* from certain countries, which makes these countries dominate the overall results for all the enterprises. To deal with this issue, we investigate the countries that dominate each enterprise's *Persistent Communities.* In particular, for each enterprise, we rank the countries based on the number of *Persistent Hosts* they include, then add the ranks of the same country for all the enterprises, and finally calculate the average rank. The results are shown in Table 4. The column *Appearance* is the number of enterprises having *Persistent Hosts* from that country. For instance, 82 out of 84 enterprises have *Persistent Hosts* from US.

The top countries that have most and highest ranks of *Persistent Hosts* may vary depending on many factors, like the enterprises' locations, the services they rely on and so forth. For example, most of the *Persistent Hosts* of a company located in Germany might be from Europe. However, even the top countries for *Persistent Hosts* change, they may match the countries that have the most hosts talking to the enterprises. Now we investigate whether this is true. We list the countries that have most hosts communicating with the enterprises in Table 3. From Table 2, Table 3 and Table 4 we can see that the top countries are very similar. 8 countries in Table 2 and 9 countries in Table 4 appear in Table 3. The two countries only appear in Table 2 and Table 4 but not in Table 3 are Japan and Australia, and they rank as 12th and 14th in Table 3, which are very close to top 10. Consequently, the top countries for *Persistent Hosts* are similar to the top countries for all the hosts that communicate with the enterprises.

4 Multi-Enterprise Persistent Communities

In the last section, we studied the *Persistent Community* for a single enterprise. Here we correlate the *Persistent Hosts* between multiple enterprises and characterize them from several aspects. We define a *Multi-Enterprise Persistent Host*

Table 2: Countries having Most Persistent Hosts

Rank	Country	Persistent Hosts
1	US	1493286
2	Netherlands	530606
3	UK	156392
4	Canada	108282
5	Russian Federation	81165
6	Germany	79999
7	China	75418
8	Australia	59589
9	Japan	57608
10	France	57229

Table 3: Countries having Most Hosts

Rank	Country	All Hosts
1	US	60290311
2	China	15369512
3	UK	10025163
4	Netherlands	6599877
5	Russian Federation	5996232
6	Germany	5907806
7	Brazil	5713122
8	France	5460201
9	Canada	5152875
10	India	4700297

as the host that communicates with M ($M \geq 2$) enterprises at least N days out of 7 days, and define a *Multi-Enterprise Persistent Community* as the set of these hosts. In this section, we use 7 as N for analysis.

4.1 Size of Multi-Enterprise Persistent Community

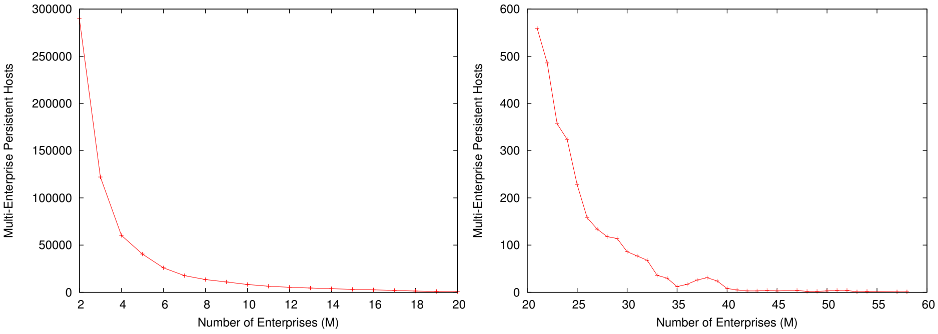
There are 3543037 *Persistent Hosts* for all the enterprises in total, and around 17.6% of them (622911) belong to *Multi-Enterprise Persistent Communities* ($M \geq 2$). We show the relationship between the number of *Multi-Enterprise Persistent Hosts* and M in Figure 5a and Figure 5b. The x-axis in the figures is M and y-axis is the number of *Multi-Enterprise Persistent Hosts*. For example, (30, 77) in Figure 5b means there are 77 hosts communicating with 30 enterprises every day in one week. From the figures we can see that as M increases, the number of *Multi-Enterprise Persistent Hosts* decreases dramatically.

4.2 Organizations of Multi-Enterprise Persistent Communities

In this part we investigate the organizations that *Multi-Enterprise Persistent Hosts* belong to. For each *Multi-Enterprise Persistent Host*, we use *MaxMind* [2] to check the organization it belongs to. The top organizations that have the most *Multi-Enterprise Persistent Hosts* are shown in Table 5. M in the first row indicates the number of enterprises (M) that *Multi-Enterprise Persistent Hosts* communicate with. For the sake of data privacy considerations, except for a handful of very large, household name organizations, this discussion strips out the specific identities of the organizations, and assigns pseudonyms. ISP 1 to ISP 10 are Internet Service Providers from US, ISP 11 to ISP 13 are from Europe

Table 4: Countries having Highest Average Rank for Persistent Hosts

Ranks	Country	Appearance	Avg Rank
1	US	82	1.54
2	UK	75	4.69
3	Netherlands	81	6.30
4	Germany	71	7.56
5	Canada	71	7.67
6	China	63	7.75
7	France	63	10.06
8	Russian Federation	56	13.16
9	Japan	65	13.43
10	Brazil	58	14.97



(a)

(b)

Fig. 5: Relationship between Number of Persistent Hosts and M

and ISP 14 to ISP 17 are from China. Univ 1 is a university located in US. Inc.1 to Inc.5 are large organizations that provide a variety of critical services to other enterprises, such as web browsers, domain service, security services, cloud services, collocation and hosting, and analytics.

The table shows that ISPs dominate the top organizations. The *Persistent Hosts* from ISPs is a complex mix, but they could be the loyal customers of an enterprise (e.g., the people who visit a bank website for transactions every day). Besides, many ISPs provide cloud storage to enterprises to keep their data safe in the event of a natural disaster or disruption in service, as well as share data with their customers. Other than ISPs, the top organizations also include some companies that provide critical services to enterprises, like Inc.1 to Inc.5. Amazon and Google also provide cloud storage and other big data services. Moreover, some of the top organizations provide services that employees rely on. For instance, employees frequently use Google or Baidu to search for information, and visit Facebook for fun. One interesting thing we noticed is that when *M=32*, a university is the organization that has most *Multi-Enterprise Persistent Hosts*. By doing an Internet search, we found that the researchers at this university did lots of network security studies and they used to host botnet ZeuS C&Cs and fake URLs. Our guess is that they may do some research work related to network security (e.g., perform probes to a large IP space frequently).

From the table we can also find that when *M=1*, 9 out of the top 10 organizations are ISPs. As M increases, within the top 10 organizations, the number of ISPs decreases and the companies providing services increases. when *M=32*, only 2 out of top 10 organizations are ISPs. This is expected because these companies are the leaders in their areas thus many enterprises use their services.

Table 5: Organizations of Multi-Enterprise Persistent Hosts

M=1	M=2	M=4	M=8	M=16	M=32
ISP1-US	ISP1-US	ISP5-US	Google	Google	Univ1
ISP2-US	ISP3-US	ISP4-US	Amazon	Microsoft	Baidu
ISP3-US	ISP5-US	ISP6-US	ISP6-US	Amazon	Google
ISP11-EU	Amazon	Amazon	Microsoft	ISP15-CH	Microsoft
ISP12-EU	ISP4-US	ISP3-US	ISP3-US	ISP14-CH	Inc.2
ISP13-EU	Google	Google	ISP14-CH	Facebook	ISP14-CH
ISP7-US	ISP6-US	Microsoft	ISP17-CH	ISP16-CH	Inc.3
Amazon	Akamai	Akamai	ISP15-CH	ISP3-US	Inc.4
ISP8-US	Microsoft	Inc.1	ISP16-CH	ISP2-US	ISP10-US
ISP9-US	ISP2-US	ISP14-CH	Akamai	Baidu	Inc.5

4.3 Multi-Enterprise Persistent Community Communication Frequency

In this part, we investigate the frequency at which *Multi-Enterprise Persistent Hosts* communicate with enterprises. In particular, we use one hour as the time window and measure how many hours per day a *Multi-Enterprise Persistent Host* communicates with enterprises.

Figure 6 shows the relationship between the number of communication hours and the number of enterprises that a *Multi-Enterprise Persistent Host* talks to (denoted as M above). From the figure we can see, in general when M increases, the number of hours also increases. This means that the more enterprises a *Multi-Enterprise Persistent Host* communicates with, the more often they have communications. One possible explanation is that these *Multi-Enterprise Persistent Hosts* provide popular services to many enterprises thus they communicate very often. In fact this explanation matches the results in Table 5, which is when

M increases, the number of companies providing services increases. We also notice that there are some outliers around (37, 10.3) and (52, 8.7). The reason is that when M is greater than 30, the number of *Multi-Enterprise Persistent Hosts* is very small, making it easy to introduce outliers.

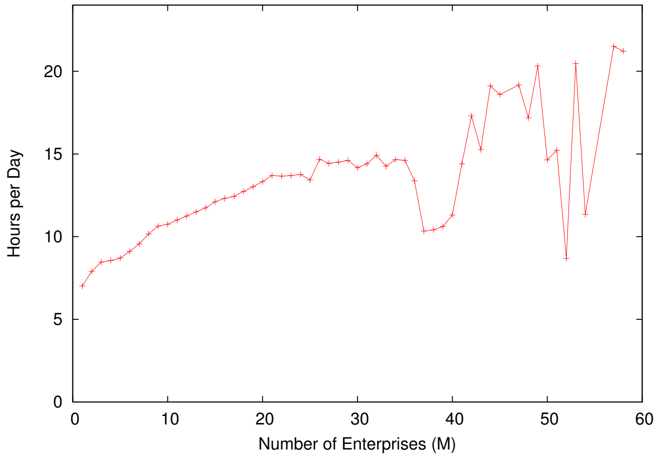


Fig. 6: Communication Hours for Multi-Enterprise Persistent Hosts

5 Do Persistent Hosts Perform DDoS Attacks?

In this section, we investigate whether the *Persistent Hosts* perform DDoS attacks by correlating them with real attack data. In particular, first we use various thresholds of traffic volume to extract attackers and investigate how many of them are *Persistent Hosts*. Then we compare the behaviors of all the *Persistent Hosts* before the attacks and during the attacks.

5.1 DDoS Attacks Analysis

Data set *DDoSAlerts* includes several types of DDoS attacks, like TCP SYN, Fragment, Bandwidth and so forth. Here we focus on the bandwidth attacks. We extract 9 attacks occurring in March. Some information about these attacks is listed as follows and also in Table 6.

1. Attack 1: Among all the packets sent to the victim enterprise, more than 99.9% is TCP, 97.7% is sent to victims' port 80 from 4953 hosts, 81.7% is sent to a single victim's port 80
2. Attack 2: Among all the packets sent to the victim enterprise, more than 99.9% is TCP, 99.2% is sent to victims' port 80 from 171302 hosts. 98% is sent to a single victim's port 80

3. Attack 3: Among all the packets sent to the victim enterprise, 66.7% is UDP and 30% is TCP, 66.5% is sent to victim A's ports ranging from 1 to 65535 via UDP, 18.4% traffic is sent to victim B's ports ranging from 1 to 65535 via TCP
4. Attack 4: Among all the packets sent to the victim enterprise, 91.6% is TCP and 8.4% is UDP. 87% is sent to a single victim's ports raging from 1 to 65535 via TCP and UDP
5. Attack 5: Among all the packets sent to the victim enterprise, 99.1% is sent to DNS port 53 from 16514 hosts. 49.8% is sent to victim A's DNS port 53, 49.2% is sent to victim B's DNS port 53
6. Attack 6: Among all the packets sent to the victim enterprise, 91% is sent to DNS port 53 from 260626 hosts. 45.8% is sent to victim A's DNS port 53, 45.2% is sent to victim B's DNS port 53
7. Attack 7: Among all the packets sent to the victim enterprise, 99.7% is sent to DNS port 53 from 3738 hosts. 50% is sent to victim A's DNS port 53, 49.6% is sent to victim B's DNS port 53
8. Attack 8: Among all the packets sent to the victim enterprise, 99.7% is sent to DNS port 53 from 56718 hosts. 50% is sent to victim A's DNS port 53, 49.5% is sent to victim A's DNS port 53
9. Attack 9: Among all the packets sent to the victim enterprise, 99.3% is sent to DNS port 53 from 25655 hosts. 49.6% is sent to victim B's DNS port 53, 49.7% is sent to victim B's DNS port 53

Table 6: DDoS Attacks

Attacks	Type	Date	Start Time	End Time
Attack 1	TCP Port 80	March 19th	20:45	21:04
Attack 2	TCP Port 80	March 19th	22:21	22:34
Attack 3	TCP/UDP All Ports	March 20th	12:08	12:34
Attack 4	TCP/UDP All Ports	March 20th	13:13	13:20
Attack 5	UDP Port 53	March 25th	3:46	4:04
Attack 6	UDP Port 53	March 26th	15:27	15:34
Attack 7	UDP Port 53	March 29th	3:51	4:04
Attack 8	UDP Port 53	March 29th	11:01	11:22
Attack 9	UDP Port 53	March 30th	16:34	16:45

5.2 Identify Attackers based on Top Talkers

Each record in *DDoSAlerts* includes several fields for a single attack, including attack start/end time, victim enterprise id, and more. However, the attackers' IP addresses are not included. So the first step is to identify the attackers. In fact identifying the attackers is a hard problem, especially from the sampled data, which is outside the scope of this paper. To make the things simple, we use thresholds to extract the top talkers based on traffic volume and consider them as attackers, the procedure is shown as follows.

1. Calculate the amount of traffic that every host sends to the victim enterprise during the attack using flow records in *OneMonthFlows*
2. Calculate the average (*avg*) and standard deviation (*stddev*) of the above results
3. Calculate three thresholds as *avg+stddev*, *avg+3∗stddev* and *avg+5∗stddev*
4. Compare the volume of traffic sent from a host to the victim enterprise during the attack with the above threshold, if it is greater than the threshold, then the host is considered an attacker.

After extracting the attackers by using different thresholds, we correlate them with *Persistent Hosts* and investigate how many of *Persistent Hosts* are considered as attackers. The *Persistent Hosts* for correlation with attack 1 to attack 4 are extracted from March 12th to 18th, and the ones for attack 5 to attack 9 are extracted from March 18th to 24th. The correlation results are shown in Table 7, N_A and N_{PH} stand for the number of attackers and how many of them are *Persistent Hosts*. From the table we can see that for all the 9 attacks only 7 *Persistent Hosts* are considered as attackers, 0.14% of all the 5084 attackers. All these 7 *Persistent Hosts* are only involved in attack 8. We will discuss more about attack 8 in Section 5.4.

Table 7: Number of Attackers Using Various Thresholds

	avg+5∗stddev		avg+3∗stddev		avg+std	
	N_A	N_{PH}	N_A	N_{PH}	N_A	N_{PH}
Attack 1	30	0	106	0	476	0
Attack 2	570	0	896	0	1623	0
Attack 3	12	0	23	0	102	0
Attack 4	171	0	244	0	1337	0
Attack 5	75	0	108	0	172	0
Attack 6	89	0	119	0	312	0
Attack 7	1	0	1	0	138	0
Attack 8	682	0	770	1	776	7
Attack 9	99	0	120	0	148	0
Total	1729	0	2387	1	5084	7

5.3 Persistent Hosts' Behaviors Before Attacks and During Attacks

In the last part, we used several thresholds to identify top talkers and consider them attackers. However, there could be two issues by doing this. The first is that we estimate the hosts' traffic volume by using the packet-sampled flow records, thus we may misorder the top talkers and miss the real ones. The second is that the DDoS attacks can be performed by a great amount of attackers instead of just top talkers, thus top talkers is just a subset of all the attackers. To deal with these issues, we use another method to investigate whether the *Persistent Hosts* perform DDoS attacks in this section. In particular, we compare the behaviors of all the *Persistent Hosts* that communicate with the victim enterprise before the attack and during the attack. The time before the attack is the same time of day as the attack's, but one week earlier. For example, if the attack occurs from 22:00 to 22:30 on March 19th, we pick 22:00 to 22:30 on March 12th as the time before the attack. The reason why we select one week earlier instead of one day is because usually the traffic has weekly pattern. Like the last section, the *Persistent Hosts* for correlation with attack 1 to attack 4 are extracted from March 12th to 18th, and the ones for attack 5 to attack 9 are extracted from March 18th to 24th.

First we compare the number of *Persistent Hosts* that communicate with the victim enterprise before the attacks and during the attacks. We denote the number of *Persistent Hosts* that communicate with the victim enterprise during a certain time as N_{PH}, and denote the number of all the hosts that communicate with the victim enterprise during the same time as N_H. Then we calculate N_H and N_{PH} for the period of time before the attacks and during the attacks. The results are shown in Table 8. From the table we can find that N_H increases significantly during the attacks, while N_{PH} doesn't change greatly during the attacks. However, as with the results in the last section, attack 8 is an exception again, where the number of *Persistent Hosts* increased three times during the attack.

Besides the number of *Persistent Hosts*, we also compare the amount of traffic sent from the *Persistent Hosts* to the victims before the attacks and during the attacks. We denote the number of packets sent from *Persistent Hosts* to the victim enterprise during a certain time as P_{PH}, and denote the number of packets sent from all the hosts to the victim enterprise during the same time as P_H. We calculate P_{PH} and P_H for the period of time before the attacks and during the attacks. The results are shown in Table 9. From the table we can find that P_H during the attacks is at least hundreds of times of the P_H before the attacks. However, *Persistent Hosts* send consistent amount of traffic to the victim enterprise before the attacks and during the attacks. Again, attack 8 is an exception, where the traffic from *Persistent Hosts* increases around 22 times during the attack. The table also tells that the *Persistent Hosts* contribute greatly to the overall traffic when there is no attack, which matches the results in Section 3.4. On the contrary, the *Persistent Hosts* contribute a very small piece to the overall attack traffic.

Table 8: Number of Persistent Hosts Before Attacks and During Attacks

	Before Attacks			During Attacks		
	Date	N_H	N_{PH}	Date	N_H	N_{PH}
Attack 1	Mar 12th	53	25	Mar 19th	5357	23
Attack 2	Mar 12th	46	24	Mar 19th	172809	23
Attack 3	Mar 13th	44	22	Mar 20th	57756	24
Attack 4	Mar 13th	33	23	Mar 20th	170215	18
Attack 5	Mar 18th	3131	2087	Mar 25th	16661	2241
Attack 6	Mar 19th	3196	2032	Mar 26th	263995	3324
Attack 7	Mar 22nd	2534	1837	Mar 29th	3872	2114
Attack 8	Mar 22nd	3312	2226	Mar 29th	56803	7307
Attack 9	Mar 23rd	3210	2350	Mar 30th	26214	2055
Total	N/A	15559	10626	N/A	773682	17129

5.4 Discussion

For all the attacks except attack 8, from Table 7, Table 8 and Table 9 we can find the following things:

1. None of the *Persistent Hosts* is labeled as a top talker.
2. The number of *Persistent Hosts* that communicate with the victim enterprise before the attacks and during the attacks doesn't change greatly.
3. The volume of traffic sent from *Persistent Hosts* to the victim enterprise before the attacks and during the attacks is consistent.
4. The *Persistent Hosts* contribute a very small piece to the attack traffic, especially for the TCP-based DDoS attacks (attack 1 to 4).

Consequently, we can conclude that the *Persistent Hosts* didn't participate significantly in these 8 DDoS attacks.

However, the behaviors of the *Persistent Hosts* in attack 8 are different from the above ones. Specifically, several *Persistent Hosts* are considered as top talkers, the number of *Persistent Hosts* increases three times during the attack, and the packets they send to the victim enterprise during the attack is 22 times more than before the attack. This evidence indicates that some of these *Persistent Hosts* were involved in the attack. Although we are not able to find the reason for this attack due to the limitations of our data set, we provide three possible reasons as follows:

1. Some of the *Persistent Hosts* are public DNS resolvers and the attackers use them to perform reflection attacks. In fact, the attackers can perform very serious DDoS attacks by using public DNS servers [9], [10]. We can investigate whether these *Persistent Hosts* are DNS resolvers by using the results in other work. For example, in [23] the authors use the DNS data

Table 9: Amount of Traffic from Persistent Hosts Before Attacks and During Attacks

	Before Attacks			During Attacks		
	P_H	P_{PH}	Percent	P_H	P_{PH}	Percent
Attack 1	51733	45436	87.8%	53692878	45042	0.0%
Attack 2	65403	40176	61.4%	642544434	58101	0.0%
Attack 3	73236	62932	85.9%	1165940276	79143	0.0%
Attack 4	34458	30685	89.1%	1222748260	21195	0.0%
Attack 5	9726006	5635279	57.9%	4191548915	8155196	0.19%
Attack 6	13527414	10347669	76.5%	2223699208	15533966	0.69%
Attack 7	7684929	6056275	78.8%	1891971614	9078820	0.48%
Attack 8	11917227	8989792	75.4%	4340818177	204158593	4.70%
Attack 9	12111281	10180806	84.1%	1505631644	9924632	0.66%

taken from one of the 13 servers for the **.com/.net** registry to study the global resolvers' behaviors. We can compare their results with the above *Persistent Hosts*, and see how many of them are DNS resolvers.

2. The attackers perform attacks by spoofing many IPs and sending data to the victims. Some of these spoofed IPs belong to the *Persistent Hosts*.
3. Some of the *Persistent Hosts* are infected and perform attacks directly on the victim enterprise.

Correlating the above analysis with the fact that attacks 1 to 4 are TCP-based DDoS attacks and attacks 5 to 9 are UDP-based DNS DDoS attacks, we conclude that the *Persistent Hosts* didn't participate significantly in the TCP-based DDoS attacks, but some of them were involved in the UDP-based DDoS attacks. The low number of attacks using TCP is most likely due to the fact that TCP connection requires a three-way handshake and thus the attackers cannot spoof IPs. However, the attackers can spoof the source IPs easily as 25% of the Autonomous Systems world-wide allow IP spoofing [4]. Besides DNS, the attackers can utilize many other UDP applications to perform DDoS attacks by using spoofed addresses [24], [25].

6 Applications Based on Persistent Communities

In the earlier sections of this paper, we have investigated several characteristics of *Persistent Communities* and correlated them with real DDoS attacks. In this part, we introduce several applications that can be benefited by using *Persistent Communities.*

6.1 Improve DDoS Protection Systems

The *Persistent Community* can be used as a whitelist in DDoS detection system to reduce the amount of traffic being monitored. In Section 3.4 we have shown that *Persistent Hosts* contribute greatly to the overall traffic sent to the enterprise when there is no attack. On the contrary, they contribute very little when attacks occur, as described Section 5.3. Consequently, we can ignore the great amount of traffic sent from *Persistent Hosts* and spend more computation resources on the *Non-Persistent Hosts*. It is true that some of attack traffic could be sent from the *Persistent Hosts*, like attack 8. However, we can enhance the whitelist by considering protocol or application. For example, we can only whitelist the non-DNS, non-ICMP or only TCP traffic sent from the *Persistent Hosts*.

Notice that we are not assuming that all the *Persistent Hosts* are always benign. Instead, we claim that even some *Persistent Hosts* perform attacks, i) they don't have the ability to crash an enterprise's network without help from a large number of other attackers. The reason is that both the number of *Persistent Hosts* and the traffic they send to the victim enterprise is a tiny piece of the overall traffic during the attacks, as shown in Table 8 and Table 9. ii) we can still detect the attack without analyzing the traffic sent from the *Persistent Hosts*. The reason is that most of traffic during attacks is sent from *Non-Persistent Hosts*, as shown in Table 9.

Such a whitelist can also be used in DDoS mitigation systems to let the traffic from *Persistent Hosts* go through the enterprises during the attacks. It's true that a *Persistent Host* could either source an attack or have its IP spoofed. In this case, we let attack traffic go through. However, notice that it's one thing to cut off a *Persistent Host* of an enterprise who's sourcing attack traffic or appears to be, but it's a much bigger deal to cut off a *Persistent Host* that could be providing a business-critical service. For instance, we wouldn't want to cut off outsourced email or a payment processor even if attack traffic seems to be coming from them.

Besides using *Persistent Hosts* as a whitelist, one could apply different policies and detection algorithms between a *Persistent Community* and *Non-Persistent Community* because they have different behavior patterns. For example, one could apply conservative detection parameters and algorithms on the *Persistent Community* while using more aggressive ones on the *Non-Persistent Community* due to the fact that *Persistent Hosts* contribute very little to the attacks.

6.2 Prioritize Traffic

The *Persistent Hosts* can also be used for traffic prioritization. For example, during the period of heavy traffic, we can assign higher priority to the traffic sent from *Persistent Hosts*. Moreover, the security service providers who collect traffic information from multiple enterprises can assign different priorities to different types of *Persistent Hosts*. Recall that in section 4.2 and section 4.3 we have shown that the more enterprises a *Persistent Host* communicates with, the more often it communicates with the enterprises and the more likely it is to provide services. Based on these findings, the traffic from *Multi-Enterprise Persistent Hosts* can be assigned higher priority than *Single-Enterprise Persistent Hosts.*

6.3 Reflect Enterprise Changes in Various Aspects

We can use *Persistent Hosts* to detect enterprise changes in various aspects. For instance, a sudden decrease in *Persistent Hosts* may indicate topology changes (accidental or otherwise). As we described in Figure 1, the outlier line is a result of the reduced number of edge routers exporting NetFlow records. It's true that this change can also be detected by looking at the total number of flow records or other information. However, some other changes may not be easily identified by only considering the overall traffic. For example, a decrease in the size of a *Persistent Community* could indicate that the enterprises are losing loyal customers, or something abnormal is happening to the companies that provide services to these enterprises.

7 Related Work

Although Internet has been around for a few decades, people began to pay a attention to enterprise networks after 2000. Pang provides the first broad overview of internal enterprise traffic, including traffic makeup, patterns of locality, characteristics of various applications and so forth in [14]. In [17], the authors develop several heuristics to identify Yahoo! IP addresses, localize their locations, and then analyze D2D and client traffic characteristics. Moreover, the media conferencing traffic in global enterprise is characterized in [18].

The *Community of Interest* (COI) of an enterprise is similar to the *Persistent Hosts* in our paper. In [11], the authors introduce a methodology for evaluating various aspects of COIs of hosts within a large enterprise network. They also define the core COI based on communication frequency and popularity, then investigate their stability over time. A COI can be used to solve network security problems. PRIMED [12] generates network-wide bad COI and per-customer good COI based on history information, and then uses them together with customer-specific policies to mitigate DDoS attacks. In [13], the authors use NetFlow data to correlate the information from various organizations belonging to the same community, and then generate alerts when suspicious activity is detected. McDaniel uses COI to generate profiles for each end host and apply it to protect security threats within enterprise [16].

In our work, besides characterizing the *Persistent Community*, we also introduce some ways to use them for solving security problems, like building whitelists and prioritizing traffic. In [15], the authors introduce similar ideas. Specifically, the authors apply several heuristics to identify good COIs that have previous good communications with the enterprises and prioritize their traffic during periods of heavy traffic. However, they consider the IPs that have been seen in the past several days as benign while we focus on the hosts that appear more frequently. Another difference is that we focus on DDoS attacks while they do not. There are also some works trying to detect DDoS attacks based on the IP history. In [19], Peng utilizes nonparametric change point detection algorithm (CUSUM) based on the number of IPs that have not communicated with the protected network before, to detect DDoS attacks. In another work [20], he introduces a similar method to protect the legitimate traffic by considering whether a host has appeared before as well as how many packets have been exchanged.

8 Future Work

In our data set, part of the flow records are collected after packet sampling, which introduces two issues. The first is that we have to estimate the amount of traffic for each flow. In this paper we make a rough estimation by simply multiplying the number of packets and the packet sample ratio, but it could be more accurate. Our expectation is that the impact of packet sampling on communication between *Persistent Communities* and enterprises is less than on individual flows because we are considering a huge number of flows. The second issue of packet sampling is that we miss some flows, especially the short ones. Measuring the impacts of packet sampling on these two aspects is one of our future work.

Although we define a *Persistent Host* as the host who communicates with an enterprise at least N out of seven days, we only use 7 as N in most parts of this paper. We plan to use various values as N and investigate their results' differences.

Besides characterizing *Persistent Communities*, we also study whether they perform DDoS attacks. But we are only able to focus on the attackers related to bandwidth due to the data set limitation. Investigating application and protocol level DDoS attacks and correlating them with *Persistent Communities* is work we plan to do in the future.

9 Conclusions

The hosts that communicate with an enterprise form a complex mix, which could be composed of enterprises' customers, services provides, even scanners and DDoS attackers. Although these hosts change over time, some of them have communications with the enterprises very frequently. We define a *Persistent Host* as the host that communicates with enterprises at least N out of seven days. Measuring these *Persistent Hosts* benefits enterprises in several ways, including network management, traffic engineering, security and more. In this paper, we investigate the characteristics of *Persistent Hosts* by analyzing 78 billion flow records collected from 84 enterprises for an entire month. The enterprises cover great variety, including bank, web hosting, retails, telecommunications, and so forth, thus we expect our study is representative for many enterprises. Specifically, we characterize the *Persistent Hosts* for single enterprise from various perspectives, including their sizes, the contribution to overall traffic, country distribution and more. We also correlate the *Persistent Hosts* between multiple enterprises and characterize them from various aspects, including the their organizations, communication frequency, and so forth. Moreover, we correlate the *Persistent Hosts* with real DDoS attacks and find that some of the *Persistent Hosts* are involved in UDP-based DDoS attacks but they don't perform TCP-based attacks. Finally, we make recommendations on how to use *Persistent Hosts* for improving DDoS protection systems, prioritizing traffic and more.

References

1. VeriSign DDoS Protection: https://www.verisigninc.com/en_US/website-availability/ddos-protection/ddos-product-features/index.xhtml
2. MaxMind: http://www.maxmind.com
3. sFlow: http://www.sflow.org/
4. The Spoofer Project. http://spoofer.cmand.org.
5. NetFlow: http://www.cisco.com/c/en/us/products/ios-nx-os-software/ios-netflow
6. B. Claise, RFC 3954: Cisco systems netflow services export version 9 (2004)
7. B. Claise, B. Trammell. RFC 7012: Information Model for IP Flow Information Export (IPFIX)
8. Phaal, P., S. Panchen, and N. McKee. RFC 3176: InMon Corporation's sFlow: A Method for Monitoring Traffic in Switched and Routed Networks, 2001
9. M. E. Donner; Prolexic. https://tinyurl.com/prolexic-167gbit, May 2013.
10. M. Prince; CloudFlare, Inc. http://blog.cloudflare.com/the-ddos-that-almost-broke-the-internet, March 2013.
11. Aiello, William and Kalmanek, Charles R. and McDaniel, Patrick Drew and Sen, Subhabrata and Spatscheck, Oliver and van der Merwe, Jacobus E.: Analysis of Communities of Interest in Data Networks. In: Passive and Active Measurement (PAM), 2005
12. Verkaik, Patrick and Spatscheck, Oliver and Van der Merwe, Jacobus and Snoeren, Alex C: Primed: community-of-interest-based ddos mitigation. In: Proceedings of SIGCOMM workshop on Large-scale attack defense, 2006
13. Weigert, Stefan and Hiltunen, Matti A and Fetzer, Christof: Community-based Analysis of Netflow for Early Detection of Security Incidents. In: Proceedings of LISA, 2011
14. Pang, Ruoming and Allman, Mark and Bennett, Mike and Lee, Jason and Paxson, Vern and Tierney, Brian: A first look at modern enterprise traffic. In: Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement, 2005
15. Kalafut, Andrew J and Van der Merwe, Jacobus and Gupta, Minaxi: Communities of interest for internet traffic prioritization. In: Proceedings of INFOCOM Workshops, 2009
16. McDaniel, Patrick Drew and Sen, Subhabrata and Spatscheck, Oliver and van der Merwe, Jacobus E and Aiello, William and Kalmanek, Charles R: Enterprise Security: A Community of Interest Based Approach. In: Proceedings of Network and Distributed System Security Symposium, 2006
17. Chen, Yingying and Jain, Sourabh and Adhikari, Vijay Kumar and Zhang, Zhi-Li and Xu, Kuai: A first look at inter-data center traffic characteristics via yahoo! datasets. In: Proceedings of INFOCOM, 2011

18. Vasudevan, Vijay and Sengupta, Sudipta and Li, Jin: A first look at media conferencing traffic in the global enterprise, In: Proceedings of Passive and Active Network Measurement, 2009
19. Peng, Tao and Leckie, Christopher and Ramamohanarao, Kotagiri: Proactively detecting distributed denial of service attacks using source IP address monitoring. In: Proceedings of the Third International IFIP-TC6 Networking Conference, 2004
20. Peng, Tao and Leckie, Christopher and Ramamohanarao, Kotagiri: Protection from distributed denial of service attacks using history-based IP filtering. In: Proceedings of International Conference on Communications, 2003.
21. Barakat, Chadi and Iannaccone, Gianluca and Diot, Christophe: Ranking flows from sampled traffic. In: Proceedings of the 2005 ACM conference on Emerging network experiment and technology, 2005
22. Haddadi, Hamed and Landa, Raul and Moore, Andrew W and Bhatti, Saleem and Rio, Miguel and Che, Xianhui: Revisiting the issues on netflow sample and export performance. In: Proceedings of Communications and Networking in China, 2008
23. Osterweil, Eric and McPherson, Danny and DiBenedetto, Steve and Papadopoulos, Christos and Massey, Dan: Behavior of DNSTop Talkers, a. com/. net View. In: Proceedings of Passive and Active Measurement, pp. 211-220. Springer Berlin Heidelberg, 2012
24. Christian Rossow: Amplification Hell: Revisiting Network Protocols for DDoS Abuse. Network and Distributed System Security Symposium (NDSS), 2014
25. Czyz, Jakub and Kallitsis, Michael and Gharaibeh, Manaf and Papadopoulos, Christos and Bailey, Michael and Karir, Manish: Taming the 800 Pound Gorilla: The Rise and Decline of NTP DDoS Attacks, IMC, 2014