Assignment 6: Analyze Darknet Traffic (30pts)

Due on Wednesday March 22, 2023 11:59pm

1 Introduction

The goal of this assignment is to provide hands-on experience with analyzing darknet traffic collected by UCSD's network telescope (aka *darknet*). You will write code to characterize scanning behavior and Internet backscatter in the traffic.

1.1 Data format

The dataset (`https://catalog.caida.org/dataset/annotated_anonymized_telescope_packets_sampler`) consists of two directories (`pcapsanon` and `metadata`) containing packet captures (pcaps) and packet meta-data, respectively. Both directories have 168 (=7 days x 24 hours) files, named using the regular expression `ucsd-nt\.\d{10}`. The 10-digit part in the file name is the Unix timestamp of the start time of the data.

Each pcap file in the `pcapsanon` folder contains one hour of traffic data sent toward one of the /24 prefixes monitored by the telescope. The source IP addresses have been anonymized using Crypto-PAn [1] implemented in *traceanon* [2]. The first three octets of the destination IP addresses have been re-mapped to 10.0.0.0/24, but the last octet is preserved.

The JSON files in the `metadata` directory provide meta-data for the corresponding packet traces by matching the file names in the `pcapsanon` directory. Each JSON file is structured as an array of JSON objects as shown in the following example.

```
[{
  "PacketCnt": 1,
  "IsZmap": false,
  "IsMasscan": false,
  "IsMirai": false,
  "IsBogon": false,
  "SrcASN": "24560",
  "NetacqCountry": "in",
  "MaxmindCountry": "IN",
  "KnownScanner": ""
}, ...]
```

The length of the array is equal to the total number of packets in the traffic trace. The `PacketCnt` field in each JSON object indicates the position of the packet in the corresponding trace. Other fields provide additional information about the packet. Table 1 provides a

Table 1: Description of fields in the JSON object.

Field names	Type	Description
PacketCnt	Integer	Packet ID in the corresponding pcap file
IsZmap	boolean	Does the packet match the signature of ZMap?
IsMasscan	boolean	Does the packet match the signature of MASSCAN?
IsMirai	boolean	Does the packet match the signature of Mirai?
IsBogon	boolean	Does the packet have a Bogon source IP address?
SrcASN	string	ASN that the original source IP address belong to, according to prefix-to-asn dataset [4] published on that day
NetacqCountry	string	Geolocation of the original source IP address, according to Netacq-Edge (country level)
MaxmindCountry	string	Geolocation of the original source IP address, according to Maxmind (country level)
KnownScanner	string	The name of known scanners inferred using the original source IP addresses

detailed description of the fields. Noted that the ASNs and geolocation information were resolved using the original source IP addresses. Please refer to the source code that generates the JSON file (`https://github.com/CAIDA/greynetanalysis`) for the heuristics used to identify the types of scanners.

1.2 Reading PCAP files

Pcap is the de-facto standard for storing network packets. You can use Wireshark [8] to display the content of pcap files. Online tutorials (e.g., [9]) explain how to install and use Wireshark. In this assignment, you are going to write scripts/programs to characterize dark-net traffic using different fields in packet headers. Most common programming languages have libraries (e.g., scapy [10] in Python, GoPacket [11] in Golang) to read packets from pcap files. Alternatively, you can use command-line tools (e.g., `tshark`[12]) to output packet header fields into text format for analysis [13].

1.3 Using scapy in Python

The following example code reads a `.pcap` file and parse the packets one at a time.

```
scapy_cap = rdpcap('telescopesampleanon/pcapsanon/ucsd-nt.1660435200.pcap')
for packet in scapy_cap:
    # print the size of packet
    print(packet['IP'].len)

    # print UDP destination port
    if 'UDP' in packet:
        print(packet['UDP'].dport)

    # print TCP destination port
    if 'TCP' in packet:
```

```
        print(packet['TCP'].dport)

    # print ICMP code
    if 'ICMP' in packet:
        print(packet['ICMP'].code, length)
```

1.4 Due Date

Wednesday March 22, 2023 11:59pm

1.5 Submission Instructions

You will answer the questions in a written PDF report and submit to Gradescope.

2 Tasks

Implement your program to parse the dataset to answer the following questions.

2.1 Analyze the composition of darknet traffic. (8pts)

Classify the packets according to their protocols and destination ports (for TCP/UDP traffic)/type and code (for ICMP traffic).

Produce a table (4pts) (Table 2 shows an example) reporting the percentage of traffic byte volume (using the length field in the IP headers to count bytes in a packet) and number of packets of the top five transport protocols sorted by traffic (byte) volume.

Note that you may find encapsulated packets in the traces (i.e., an IP packet as a payload of an IP packet). You only need to consider the protocol of the first IP header.

Write a paragraph (4pts) to discuss the following four questions:

1. Which *protocol, destination port* pair had the highest packet count in darknet traffic?
2. Which *protocol, destination port* pair had the highest byte volume in darknet traffic?
3. What are the three most common destination TCP and UDP ports? Which applications use those TCP/UDP ports according to the IANA registry [3]? (Look for "Service Name" in the IANA table.)
4. Why are scanners targeting these services?

Table 2: Example table for presenting results in question 1.

Rank	Protocol (Port)	Volume in Bytes (percentage)	Number of packets (percentage)
1	UDP (53)	V1 (v1%)	C1 (c1%)
2	TCP (23)	V2 (v2%)	C2 (c2%)
3	...		
4	...		
5	...		

2.2 Analyze the origin of darknet traffic. (8pts)

Rank by traffic (byte) volume the source country and autonomous system (AS), using geolocation and the AS of the source IP address (in the meta-data file).

Produce two tables (2pts) that show the top 10 *(I) countries* and *(II) ASes* (sorted by byte volume) originating darknet traffic to this /24 prefix. Use the the AS Organizations dataset linked below to resolve the countries and organizations of the top ASes.

AS Organizations: `https://cseweb.ucsd.edu/classes/wi23/cse291-e/pa4/20230101.as-org2info.txt.gz`

Plot a time-series graph (2pts) to show the change in traffic byte volume per hour from the top 10 countries.

Write a paragraph (4pts) to summarize the results and answer the following questions:

1. What are the similarities and dissimilarities between the lists of countries in the two tables?

2. What pattern did you observe in the time-series plots? Suggest possible explanations for the patterns.

3. Inspect the traffic manually, and suggest a plausible explanation for the AS that sent the most traffic to this subnet of the darknet.

4. Why might the country/origin AS information be unreliable in this traffic data?

2.3 Identify the type of scanners. (10pts)

The dataset identified **three** common Internet scanner patterns (ZMap, MASSCAN, and Mirai) and **eight** Internet scanning campaigns. You will find the eight campaigns in the metadata files under the "KnownScanner" field.

Produce a table (5pts) that show the traffic (byte) volumes sent using these 11 categories. Analyze the number of unique (*protocol, port*) pairs that each measurement campaign targeted, and the top 3 destination ports that each scanner implementation probed.

Write a paragraph (5pts) to summarize the findings and answer the following questions:

1. Of the eight Internet scanning campaigns, which one had the largest coverage in terms of 1) TCP destination ports, and 2) UDP destination ports? Which protocol (TCP and UDP) had higher coverage, and why?

2. Rapid7 performs Internet-wide TCP and UDP scans to selected destination ports [6]. Can you find the traffic in the dataset related to those two scans? If not, suggest possible explanations.

3. Why did Mirai mostly scan only a few ports? Which network services are usually hosted behind these ports?

2.4 Inferring DDoS attacks using Internet Backscatter (4pts)

TCP SYN/ACK and TCP RST packets are sometimes called *Internet Backscatter traffic*, a class of darknet traffic induced by randomly-spoofed-source denial-of-service (DDoS) attacks. In this case the source of the packet is the victim responding to the randomly spoofed source address, some of which are in our darknet prefix.

Produce a table (2pts) that shows the top 10 such victim ASes observed in the data. Use AS2Org [7] to identify the the organizations that operate those ASes.

Write a paragraph (2pts) to summarize the results and provide a possible reason that attackers targeted these ASes.

References

[1] Jun Xu, Jinliang Fan, Mostafa H. Ammar, and Sue B. Moon. "On the Design and Performance of Prefix-Preserving IP Traffic Trace Anonymization". Proceedings of ACM SIGCOMM Workshop on Internet Measurement, 2001.

[2] Shane Alcock, "Libtrace - traceanon", `https://github.com/LibtraceTeam/libtrace/wiki/traceanon`.

[3] Internet Assigned Numbers Authority (IANA), "Service Name and Transport Protocol Port Number Registry",`https://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.xhtml`

[4] CAIDA, "RouteViews IPv4 Prefix to AS mappings", `https://catalog.caida.org/dataset/routeviews_ipv4_prefix2as`

[5] CAIDA, "ASRank", `https://asrank.caida.org`

[6] Rapid7, "Open data", `https://opendata.rapid7.com`

[7] CAIDA, "AS to organizations mappings", `https://catalog.caida.org/dataset/as_organizations`

[8] Wireshark, "Wireshark - Go Deep", `https://www.wireshark.org`

[9] Scott Orgera, "How to Use Wireshark: A Complete Tutorial", `https://www.lifewire.com/wireshark-tutorial-4143298`

[10] Scapy, "Scapy: the Python-based interactive packet manipulation program & library.", `https://github.com/secdev/scapy`

[11] Google, "GoPacket", `https://github.com/google/gopacket`

[12] Wireshark, "tshark Manual page", `https://www.wireshark.org/docs/man-pages/tshark.html`

[13] newspaint, "Selecting Fields to Display in TShark", `https://newspaint.wordpress.com/2021/01/18/selecting-fields-to-display-in-tshark/`