Processing Network Capture Data

In order to solidify our understanding of the "enveloping" nature of the layered approach to networking, we're going to parse a binary dump of network data. The learning objectives for this exercise are:

- Parse binary data according to some provided specifications.
- Understand exactly how the network layers encapsulate data provided by "higher" layers, and how they handle data from lower layers.
- Play and experiment with very real network data.

By the time you are finished you will have written a program that recreates some critical sections of a small subset of network protocols including: Ethernet, IP, HTTP.

Stage Zero: What Do We Know About The Provided Data?

In order to properly parse any data, we have to know what the format of that data is. The file we've provided to you represents data in the format that it was sent across the internet. The data was collected using a network capture tool, and represents the total data sent during a single HTTP request/response cycle. The HTTP request was for a particular `jpg` image which was delivered by a server. Your ultimate goal is to extract this image data and write it to a file on your computer.

As we have learned, as data travels down the network layer hierarchy, the layer below accepts the data and wraps it with it's own format. In our capture we have 4 of the 5 layers of data represented:

- Application: HTTP
- Transport: TCP or UDP (you'll find out)
- Network: IP (v4 or v6? you'll find out)
- Link: Ethernet (what type?)

In addition to our knowledge of the internet layers, we know that this data is saved in a specific format, which is documented here. There are two notable facets to the pcap savefile: the **global header** and the **per file header**.

Even the most careful reading of the documentation will not tell you anything about which protocols are represented at each network layer. In fact, without reading the data **at each layer** all we can know is that we have data like this:

- One global header, followed by
- Some number of packets, each of which with a header that will tell you long that particular packet is.

Here are some useful facts about **this specific capture** which can help you validate your findings as you go:

- There are 99 packets.
- None of the packets have been truncated at all by the capture process
  - Meaning 100% of the data sent between the two hosts represented here
- All of the IP datagrams use the same version of IP (one of IPv4 or IPv6)
- A lossy connection was simulated, so individual packets may not have arrived exactly once, and may or may not have arrived in the order they were sent.
  - However, we do ensure that all of the data for both the HTTP request and response are represented in the capture.

> Some Advice: When doing binary parsing, it is incredibly helpful to program defensively by using lots of assertions to help us detect if something is wrong with our assumptions.

Stage One: Read The Pcap Headers

The Global Header

Before we start programming we are going to manually parse the global Pcap header. This will give us some useful practice "thinking in binary" and it will force us to encounter and tackle the concept of "endian-ness". Use the `xxd` command to turn the provided binary dump into a hex dump then, Using the pcap documentation:

- Examine the "Magic Number",
  - What order is it in?
  - What does that tell you about parsing the rest of the pcap provided data?
    - **It plays a crucial role, so don't take this step lightly**
- What are the Major and Minor versions?
- Verify that the values which are always zero are in fact zero.
- What is the snapshot length?
  - Sanity check, the snapshot length is the longest a single packet can be in the capture. Does the number look small enough to be the size of a single packet (in bytes)?
- What is the link layer header type?
  - **We will need the specification details of this header in stage 2**

The Per Packet Header

Now, we should start writing a program. The reference solution we have is in Python 3, but you can use any language you want. Your goal is to read every individual pcap packet header. You should write a program that can:

- Read the captured and total length of each packet
  - Verify that for every packet the captured length and total length are equal.
  - (this is true for the file we provided, but not in general)
  - This will also tell you where the next packet header starts.
- Verify that there are 99 Packets represented in this data.
  - (Again, this is true in the provided data, not generally)

Stage Two: Read The Ethernet Headers

You should have been able to determine from the global header the type of the Ethernet Header. It's valuable practice to first try and use Google to find the exact specification for this header.

> If you spend more than 15 minutes trying to track it down and only encounter frustration, look ahead at the **spoilers** section for a direct link to the header.

Once you have determined the format of the header, extend your program to:

- Determine the version of the wrapped IP datagram (IPv6 or IPv4) so we can parse that data.
  - Verify that all the IP datagrams have the same format.
    - (Again, true in this data, but not in an arbitrary capture)
  - Print the source and destination MAC addresses
    - Verify that the MAC addresses make sense.

Stage Three: Read The IP Headers

Once again, you should strive to find the specification of the IP header yourself, but it is linked below in the spoilers section. Once you've determined the format of the header you should be able to extend your program to:

- Verify that the version is what we expect it to be
- Determine if any of the optional headers were sent for each datagram
- Determine the length of the IP header for each datagram
  - **You'll need this information to later, so save it somehow**
- Determine the source and destination IP addresses.
  - Verify that the IP addresses make sense
- Determine the relative starting location of the enclosed transport layer segment
- Determine the transport protocol being used
  - Verify that the same transport protocol used for all the IP datagrams

Stage Four: Read The Transport Headers

Once you've parsed the IP headers, you'll know where the transport headers start, and which protocol is being used. Once again, you may find the specification yourself, and it is linked in the spoilers section. Once you have this information you should extend your program to:

- Determine the ports used to communicate
  - Sanity check: How many ports are used? Does that sound right for our transport protocol?
- Determine the length of each transport header
- Using data from this and previous layers determine where the HTTP data resides for each packet.
- Determine the sequence number for this packet
- Extract the HTTP data from the packet and store it somewhere

Stage Five: Parse The HTTP Data

You should have already stored all the data somewhere, now you need to put it in order, and parse it as an HTTP request. Extend your code to:

- Order the received data by TCP sequence number.
- Combine it into a single binary string
- Interpret that binary string as text
- Print that text and verify that the headers all make sense
- Extract just the body, write it to disk as a file and save that file with a `.jpg` extension so your operating system knows what to do with it.
- View the file.
- Rejoice

Spoilers:

- Ethernet Header Format
- IP Header Format
- Transport Layer Header Format