Title of Final Project: **Big Marketing Network Operational Maintenance**

Section: 52740

Group Number: 10

Student Name: Haoliang Hu

Student UT EID: hh27683

Student Name: Alex Hoang

Student UT EID: agh2398

Student Name: Kelly To

Student UT EID: kt24854

Student Name: Daniel Miao

Student UT EID: dm52663

Date: 11/01/2023

Analysis

Goal (or Thesis)

The project aims to delve into the activities occurring on Big Marketing's networks over a two-week period, with a focus on proactive monitoring and analysis of any network events. The investigation will employ the scientific method, using observations from the data to form a hypothesis on patterns and anomalies occurring within Big Marketing's network traffic. The end goal is to create a narrative that provides accurate and valuable insights for understanding network behavior, identifying potential issues, and informing future network management strategies.

Description of the Dataset

There are seven CSV files in total. For Week 1 data, there are four files: three chunk files of network flow data and one file for network health and status data (Big Brother data). For

Week 2 data, there are three files: one file for network flow data, one file for network health and status data (Big Brother data), and one file for Intrusion Protection System data. We decided to combine the three Week 1 chunk files for network flow data into one data frame to have the data in one place, like how it is for Week 2. For this Week 1 network flow dataset (week1_nf), there are 46,138,310 rows and 19 columns. We decided to dive deeper into the Week 1 network flow dataset specifically because, based on our initial hypothesis, network flow data was the most consequential type of data to paint a more comprehensive picture of how abnormal data might be introduced into the system. We believed that potential problems would be most evidently shown by examining any anomalies in traffic specifically, as it is a uniform measure that universally encompasses all users, contexts, settings, and events.

There are 19 column names in the dataset: TimeSeconds, parsedDate, dateTimeStr, ipLayerProtocol, ipLayerProtocolCode, firstSeenSrcIp, firstSeenDestIp, firstSeenSrcPort, firstSeenDestPort, moreFragments, contFragments, durationSeconds, firstSeenSrcPayloadBytes, firstSeenDestPayloadBytes, firstSeenSrcTotalBytes, firstSeenDestTotalBytes, firstSeenSrcPacketCount, firstSeenDestPacketCount, and recordForceOut.

The columns classified as categorical are ipLayerProtocol, ipLayerProtocolCode, firstSeenSrcIp, firstSeenDestIp, firstSeenSrcPort, and firstSeenDestPort. The column ‘ipLayerProtocol’ has 3 unique values with a mode of 6. The column ‘ipLayerProtocolCode’ has 3 unique values with a mode of 45845003. The column ‘firstSeenSrcIp’ has 1139 unique values with a mode of 9069934. The column ‘firstSeenDestIp’ has 1242 unique values with a mode of 20462043. The column ‘firstSeenSrcPort’ has 64495 unique values with a mode of 80. The column ‘firstSeenDestPort’ has 65536 unique values with a mode of 80.

The columns classified as ordinal are TimeSeconds, parsedDate, dateTimeStr, moreFragments, contFragments, durationSeconds, firstSeenSrcPayloadBytes, firstSeenDestPayloadBytes, firstSeenSrcTotalBytes, firstSeenDestTotalBytes, firstSeenSrcPacketCount, firstSeenDestPacketCount, and recordForceOut. Out of the ordinal columns, the columns classified as binary are moreFragments, contFragments, and recordForceOut. The columns classified as discrete are durationSeconds, firstSeenSrcPayloadBytes, firstSeenDestPayloadBytes, firstSeenSrcTotalBytes, firstSeenDestTotalBytes, firstSeenSrcPacketCount, and firstSeenDestPacketCount. The column durationSeconds is measured in seconds, and has a range of 1800, median of 3, mean of 1.998480e01, and standard deviation of 62.6364. The column firstSeenSrcPayloadBytes is measured in bytes, and has a range of 6.109244e08, median of 19, mean of 33.66539, and standard deviation of 95705.42. The column firstSeenDestPayloadBytes has a range of 1.070089e07, median of 503, mean of 20045.48, and standard deviation of 3.076807e05. The column firstSeenSrcTotalBytes has a range of 6.348302e08, median of 297, mean of 1094.964, and standard deviation of 99782.63. The column firstSeenDestTotalBytes has a range of 1.220066e07, median of 619, mean of 20965.80, and standard deviation of 319869.8. The column firstSeenSrcPacketCount has a range of 442699, median of 5, mean of 13.61452, and standard deviation of 150.3247. The column firstSeenDestPacketCount has a range of 2.259180e05, median of 2, mean of 16.89193, and standard deviation of 229.9829. The columns classified as continuous are TimeSeconds, parsedDate, and dateTimeStr. For the continuous data, namely variables TimeSeconds, the units are time in seconds. dateTimeStr and parsedDate do not make sense to find the statistics for because the numbers all conform to a unique and specific categorical integer format. TimeSeconds has a range of 5.23e05, median of 1.364984e09, mean of 1.364990e09, and standard deviation of 1.109118e05.

Data Wrangling

For the Week 1 network flow dataset (week1_nf), there are no missing values, so no replacement of missing data was necessary. Additionally, no rows were deleted, as we did not have any rows with completely empty values or extraneous/duplicated rows. The data was all in the correct range because most of the data falls within the correct range of types of data measured by the IP network. The amount of time and data (bytes) transmitted can vary widely given the needs of the network and the types of downloads/uploads that were being utilized. However, there were no negative values or other values that were too large in the context of the attribute being measured. We changed the datatype of the column parsedDate from object to datetime to be able to manipulate the values as dates and times. We changed the datatype of the columns ipLayerProtocol, firstSeenSrcPort, and firstSeenDestPort from int64 to object since they are categorical variables despite the values being numerical. After checking for duplicates, we did not find any extraneous values.

Preliminary Statistical Analysis

To perform correlation analysis, we obtained the correlation and p-value for between attributes using the “pearsonr function” from the scipy.stats module. We filtered the results to not include any correlation that has an absolute value of 1 or below 0.3. We filtered based on these values to enhance the quality and interpretability of the results. A correlation coefficient of 1 indicates a perfect linear relationship, rendering one variable redundant, while correlations below 0.3 signify weak relationships.

Among quantitative values, the attributes that described similar data tended to have high correlations. For example, moreFragments had a high correlation with contFragments because both of the attributes described the flow record as part of a long-running data stream with marginal differences given the order in the flow as the first or a subsequent record in the same flow. Additionally, the four attributes that measure bytes are all high correlation, given they measure roughly the same items - source, destination, source total, and destination total bytes all measure address packets using different formats and the related source and destinations. However, the correlations are not noteworthy because they have similar traits - the values being compared against each other are similar, to begin with, given the unique nature of this dataset as more of a historical timeline of interconnected logs rather than a holistic dataset measuring across different contexts of attributes. Any further analysis would lead to inconsequential findings because of how related the attributes are to each other. Additionally, there is little to no relationship across different types of attributes, which is the reason why the correlation exercise is typically performed.

There was no determination of best fit because the variables do not exhibit correlations with each other outside of those with similar traits measured (network data size, fragments, or duration of time). Among those variables with similar traits measured, the relationship is linear since each attribute is measuring the same order of data, and relationships in the context of size are linear. Between most attributes, correlations were below 0.2, with most correlations between 0.13 - 0.05. Notably, the correlation between firstSeenDestPayloadBytes and firstSeenDestPacketCount / firstSeenSrcPacketCount is higher than other combinations of packet and byte attributes at 0.99 / 0.92 respectively, indicating a high degree of correlation between destination payload bytes and packet count from either the source or the destination.

Narration

- We are working for Big Marketing, an international marketing company that is helping Total Crop Protection Services roll out a marketing campaign for “Butterfly 2.0”, an altered butterfly that will lead to the extinction of natural butterflies.
- A group calling themselves the “Butterfly Warriors” has sent the firm a letter threatening to attack the firm’s network in retaliation to the marketing campaign.
- We believe that they are planning and executing malicious attacks, specifically DOS/DDOS attacks, malicious code implantation, and FTP exfiltration.
    - A sudden increase in traffic, leading to server crashes for www.bigmkt2.com but not for www.bigmkt3.com, observed in netflow data
    - Unusual FTP traffic patterns, large volume of data transferred outside Big Marketing’s network, observed in netflow data
- By analyzing the attacking IP addresses, the attacks are coordinated and there are even corporate machines that have been used by the attackers.
    - The service attacks were traced to at least 10 different IP addresses
    - Port scans occur against Site 3 from one attacker through 10.6.6.6
- Before the second week, Big Marketing managed to add an Intrusion Protection System (IPS) and successfully blocked the same type of DDos and FTP exfiltration attack techniques used in week 1.
    - This can be seen with a gap in data collection for three days followed by a significant decrease in malicious activities, as shown by IPS logs
- However, “Butterfly Warriors” started to attack in other ways, like publicizing exfiltrated customer info while implanting more malware for botnets (allowing malicious attacks on other companies).
    - Marketing representatives visit the malicious website, leading to infections. Short session durations and unusual outbound traffic patterns may be observed in netflow data
- Course of action:
    - Determine if this is an insider threat
    - Determine solutions for remediation/prevention of similar breaches in the future
    - Determine the severity and impact of data loss on Big Marketing