Network Security Project 2 Report

1. What model or algorithm you use?

I used rule-based method because after I manually inspect all the data, I found that in “Execution process ID” in “Sysmon.xml” and “Correlation Activity ID” in “Security.xml” of different person are different and in “Example Test”, further more, “Execution process ID” and “Correlation Activity ID” are identical if the data belongs to the same person.

According to above 2 discovered features, I used a dictionary to store the mappings of “Execution process ID”, “Correlation Activity ID” and “Person 1-6” so that every time a new data comes in, I only have to extract its “Execution process ID”, “Correlation Activity ID” to determine which person owns these data.

If above 2 feature can’t determine which person owns the data, I will calculate the TFIDF of urls in training data’s wireshark log as a weight vector then multiply test data’s url frequency with the weight vector and pick the highest one as the answer.

1. Anything interesting you find or problems you encounter in the whole process?

The interesting things is that at the beginning I thought that this will be hard to choose which feature to use, but there’s a bug or something that “Execution process ID” and “Correlation Activity ID” themselves are enough for me to decide these data belongs to which person.

Problems I encountered is that in real world “Execution process ID” and “Correlation Activity ID” will change every time when the computer restart, so this method certainly can not adapt to the real world scenario and can only fit in this particular situation.