COMP0060, Coursework

1. The similarity graph generated by using the import address table (IAT) similarity metrics is more connected than the one generated by using strings only, as the graph using IAT has 712 edges while the graph using strings has 316. The IAT graph has more and larger groups than the string graph as the majority of the groups from the standard similarity graph has gained more members. This is likely because malware samples often import the same DDLs and functions which are then used to perform various actions, while the strings of a file often include error messages, debug strings, configuration options and other strings that are not as common between different malware samples.

2. The values chosen for the n-grams were 3, 7 and 11 as suggested in the question description. The amount of edges they had were 332, 271 and 246 respectively. As both using n-grams and printable strings relies on braking binary files into smaller units of data, it is expected that the numbers of edges in the graph are in a similar range of values, as many binaries would share similar pieces of data. As seen by the number of connected edges, they decrease based on an increase in n-gram size.

   This is because when using smaller n-gram sizes, less detail is captured in the shorter sequences of instructions which is less likely to be unique to the specific binary file, causing more similarities to occur. On the other hand, using larger n-gram sizes enables capturing more detailed in the longer sequences of instructions which is more likely to be unique to the specific binary file, causing less similarities to occur. It is important to choose the correct n-gram size when comparing binary files, as a too small n-gram size could miss out on capturing larger patterns or techniques used in different sequences of instructions, while a too large n-gram size could cause overfitting with too much noise or variability in the data which makes it harder to to detect similarities.

3. Running the code using the Sørensen–Dice coefficient without changing any parameters results in 438 connections, which is more than the standard jaccard index of 316. It groups several malware with different names, which means the threshold needs to be adjusted. Testing several different threshold values, aiming for a graph and a number of connections close the the original graph, the final threshold value is 0.88 which resulted in 320 connections and a graph that looks mostly like the original.

4. The similarity matrix is an alternative representation of the graph generated by *listing_5_1.py*. For this specific sample of binary files, the graph is likely a better option. Although matrices can be preferable because of their easier readability with data being represented in a tabular fashion, also has the potential to provide more detail showing of the similarity values between the binary files, and could be more efficient to create and add samples to, the share amount of samples provided makes the matrix hard to read. The only helpful data provided is that the similarity samples are fairly similar based on the colour on the plot. Therefore a graph is likely the better option of a sample size of this quantity, as its easier to read, which is what the purpose of these visualisations of data are.

5. There is a wide range of options of what to do in this hypothetical, but a possible strategy could be starting with analysing the samples to identify the features that contribute to the high similarity score and then modify it based on the discoveries. One of the broader categories changes would be to change the structure or behaviour of the code while maintaining the same functionality, such as changing instruction order, adding filler code or new control structures, changing code timing and execution order, or importing different DLLs and functions that provide the same functionality. Other changes that could be done to the file could be changing the file format to something different, use a different programming language and compiler, or design the malware for a different operating system. Doing all of these changes before testing is not advised. The recommended process would be to try some of the major changes first to see their effect on the similarity score, trying different combinations of changes until the best combination is discovered.