COMP0060
Similarity Measures Coursework

Introduction

Your goals for this coursework are to

1. work individually.
2. perform a study on the interaction between similarity measures and malware classification.
3. produce a submission classifying files into different malware families.
4. produce a two page report explaining how you conducted your experiments and justifying the decisions that you made.

You will measure similarity within sets of binaries using NCD. Proceed as follows:

- Download your assigned data set (g1.zip, g2.zip, etc) from Moodle.
- Download the classes file (fileClasses.zip) that contains representatives of each class from Moodle.
- Implement NCD.
- Use NCD to measure the similarity between the binary executables and the class files. Produce a similarity matrix for all files.
- For each of the 20 binary executables, check which class is more similar to them, using the class files as representative elements of that class.
- Download the submission file corresponding to your group (submitg1.txt, submitg2.txt, etc).
- The submission file contains a list of file names and a number next to them. These files correspond to your group files. Once you have finished the NCD analysis, choose which class file is most similar to each of your group files and write the class number next to its name. For example, if group 20 has a file `4e423ababcabfaeabab` that is more similar to `class3.bin` than to the other class files, in the corresponding line of submitg20.txt, they should write:
  `4e423ababcabfaeabab;3`

- Submit the solution in a zip file. The zip file should include the submitgX.txt file and a pdf report. The report must be short (2 pages) and include a plot of the similarity matrix, and your justifications for any decisions made during the experiments. Name the zip file with your name. Do **not** modify the submitgX.txt file name.

Submit the report by the deadline given on the 20-21 COMP0060 Moodle.

File assignments

Your file assignments are as follows:

set	students	data
1	ISTATKOV, ANAND, MAO	g1.zip
2	AMIN, STATHOPOULOU, HUANG P	g2.zip
3	TONG, GONZALEZ VASQUEZ, LIN	g3.zip
4	YE, RODRIGUES, RATNAYAKE	g4.zip
5	CHEN, ZHANG	g5.zip
6	KELLY, DE BENARDINI	g6.zip
7	WEN, NG,	g7.zip
8	HUANG Y, ATKINSON	g8.zip
9	AGARWAL, ZHENG	g9.zip
10	WANG, OOI	g10.zip
11	LI, VADLAMANI	g11.zip
12	HE, PAPAVASILIOU	g12.zip
13	ZAKHAROV, THOMAS	g13.zip
14	SOON, GAO	g14.zip
15	SERELI, HASHIM	g15.zip
16	WILLIAMS, XENOFONTOS	g16.zip
17	SEKAR, GNAP	g17.zip

The NCD Measure

The Normalised Compression Distance (NCD) is used to detect similarities among strings. It is calculated as follows:

$$NCD(X,Y) = \frac{C(XY) - \min\{C(X), C(Y)\}}{\max\{C(X), C(Y)\}}, \tag{1}$$

where X and Y are input files, $C(\cdot)$ represents the size of a file after compression and XY represents the concatenation of X and Y files.

For this coursework, you need to implement your own version of NCD. You can use any compressor you consider, however, have in mind that the compressor needs to be normal. The conditions of a normal compressor are:

1. Idempotency: $C(XX) = C(X)$
2. Monotonicity: $C(XY) \geq C(X)$
3. Symmetry: $C(XY) = C(YX)$

4. Distributivity: $C(XY) + C(Z) \leq C(XZ) + C(YZ)$

Considering X, Y, Z strings and C the file size after compression.

Although this might not be achievable in all cases, we recommend to check, at least, idempotency, monotonicity and symmetry for your chosen compressor before using NCD.

Evaluation

The evaluation will consider two elements:

- Report: Decisions (30%) and Similarity Matrix (10%) (40% of the coursework).
- Classification (60% of the coursework, each correctly classified file is a 3%).

The report will be evaluated according to its clarity, taking into account your decisions over the chosen compressor and its parameters. For the similarity matrix, compare every file in your corpus: the class and the group files. Figure 1 shows an example of a similarity matrix for 2000 files.

Any late submission will be penalized according to the UCL late submission penalties policy[1].

[1] `https://www.ucl.ac.uk/academic-manual/chapters/chapter-4-assessment-framework-taught-programmes/section-3-module-assessment#3.12`

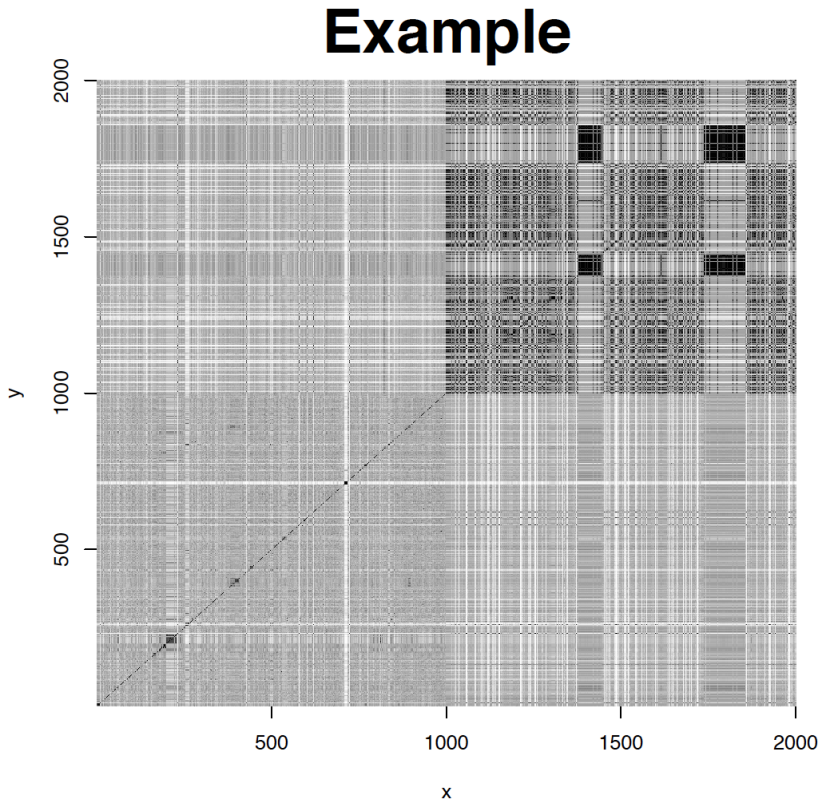

Figure 1: Example of a malware similarity matrix with 2000 files. Benign-ware files are represented at 1-1000, malware files at 1001-2000.