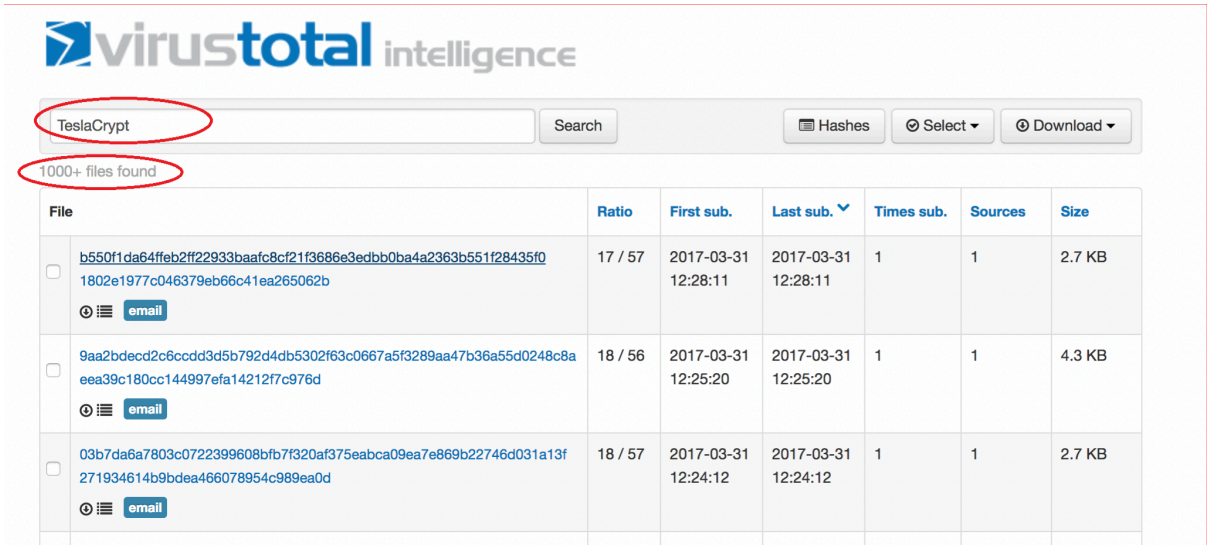
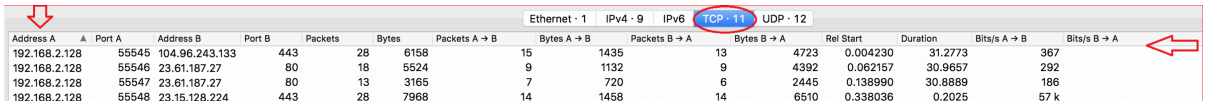
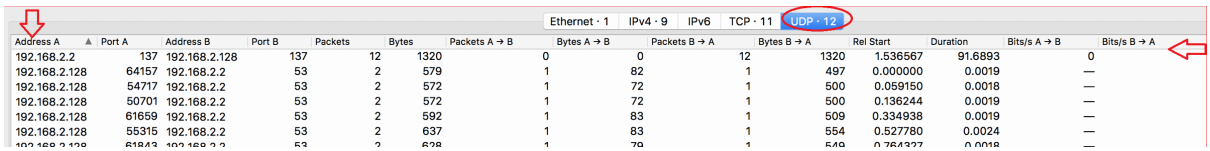
The source of Malware PCAP files

Searching for the specific ransomware families in the Virus Total Intelligence platform we could identify portable executable samples that had a corresponding behavioural analysis network traffic PCAP files. Then we filtered this information using the command lines below, to extract the conversation and related features as shown below.





Address A	Port A	Address B	Port B	Packets	Bytes	Packets A → B	Bytes A → B	Packets B → A	Bytes B → A	Rel Start	Duration	Bits/s A → B	Bits/s B → A
192.168.2.128	55545	104.96.243.133	443	28	6158	15	1435	13	4723	0.004230	31.2773	367	
192.168.2.128	55546	23.61.187.27	80	18	5524	9	1132	9	4392	0.062157	30.9657	292	
192.168.2.128	55547	23.61.187.27	80	13	3165	7	720	6	2445	0.138990	30.8889	186	
192.168.2.128	55548	23.15.128.224	443	28	7968	14	1458	14	6510	0.338036	0.2025	57 k	



Address A	Port A	Address B	Port B	Packets	Bytes	Packets A → B	Bytes A → B	Packets B → A	Bytes B → A	Rel Start	Duration	Bits/s A → B	Bits/s B → A
192.168.2.2	137	192.168.2.128	137	12	1320	0	0	12	1320	1.536567	91.6893	0	
192.168.2.128	64157	192.168.2.2	53	2	579	1	82	1	497	0.000000	0.0019	—	
192.168.2.128	54717	192.168.2.2	53	2	572	1	72	1	500	0.059150	0.0018	—	
192.168.2.128	50701	192.168.2.2	53	2	572	1	72	1	500	0.136244	0.0019	—	
192.168.2.128	61659	192.168.2.2	53	2	592	1	83	1	509	0.334938	0.0019	—	
192.168.2.128	55315	192.168.2.2	53	2	637	1	83	1	554	0.527780	0.0024	—	

Extract script

Tshark -q -z conv,tcp -z conv,udp -r "input path\input pcap" >>"output path\familyname.xls"

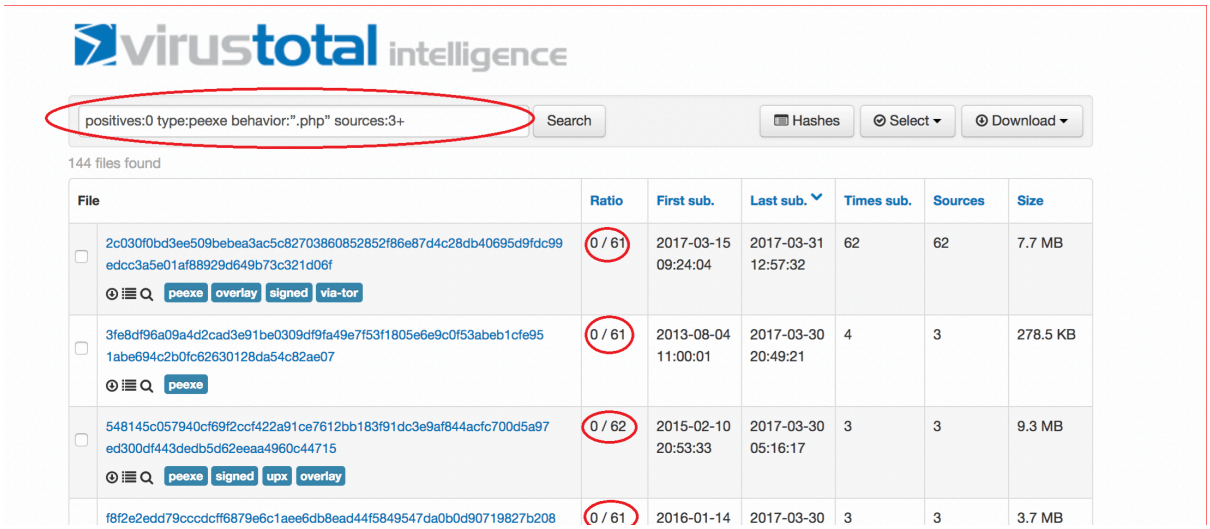
-z conv, tcp	Show TCP statistics
-z conv, udp	Show Udp statistics
-q	To print statistics
-r	Input file
>>	Append date to specified output file

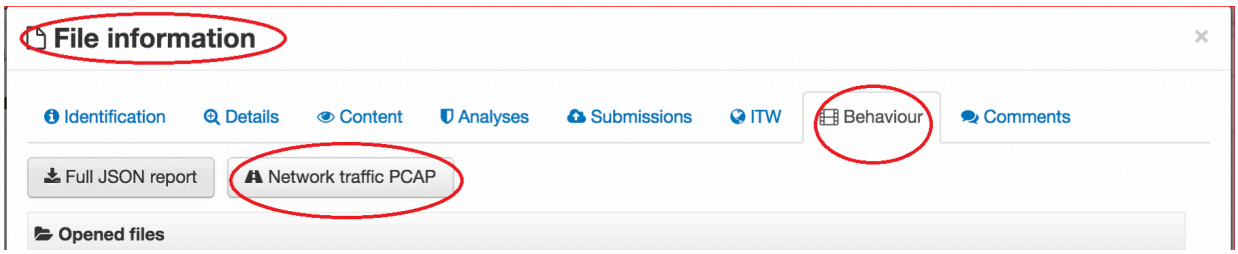
Then, the texts converted to columns using Excel. In addition, of two other columns added for our reference:

1- Protocol for Tcp/Udp conversation
2- Hash to identify which conversation belong to traffic samples.

The source of Goodware PCAP files

We used virus total threat intelligence platform, our search criteria targeted portable executable files that had been submitted at least three times and had zero detections by antivirus engines. The search was applied to several different behavioural report criteria to provide a collection of 264 goodware samples. As you can see below.





The PCAP files downloaded were filtered and features extracted using the same procedures used for malware above.

Assignment- Second Part
Filtering our Dataset

First:

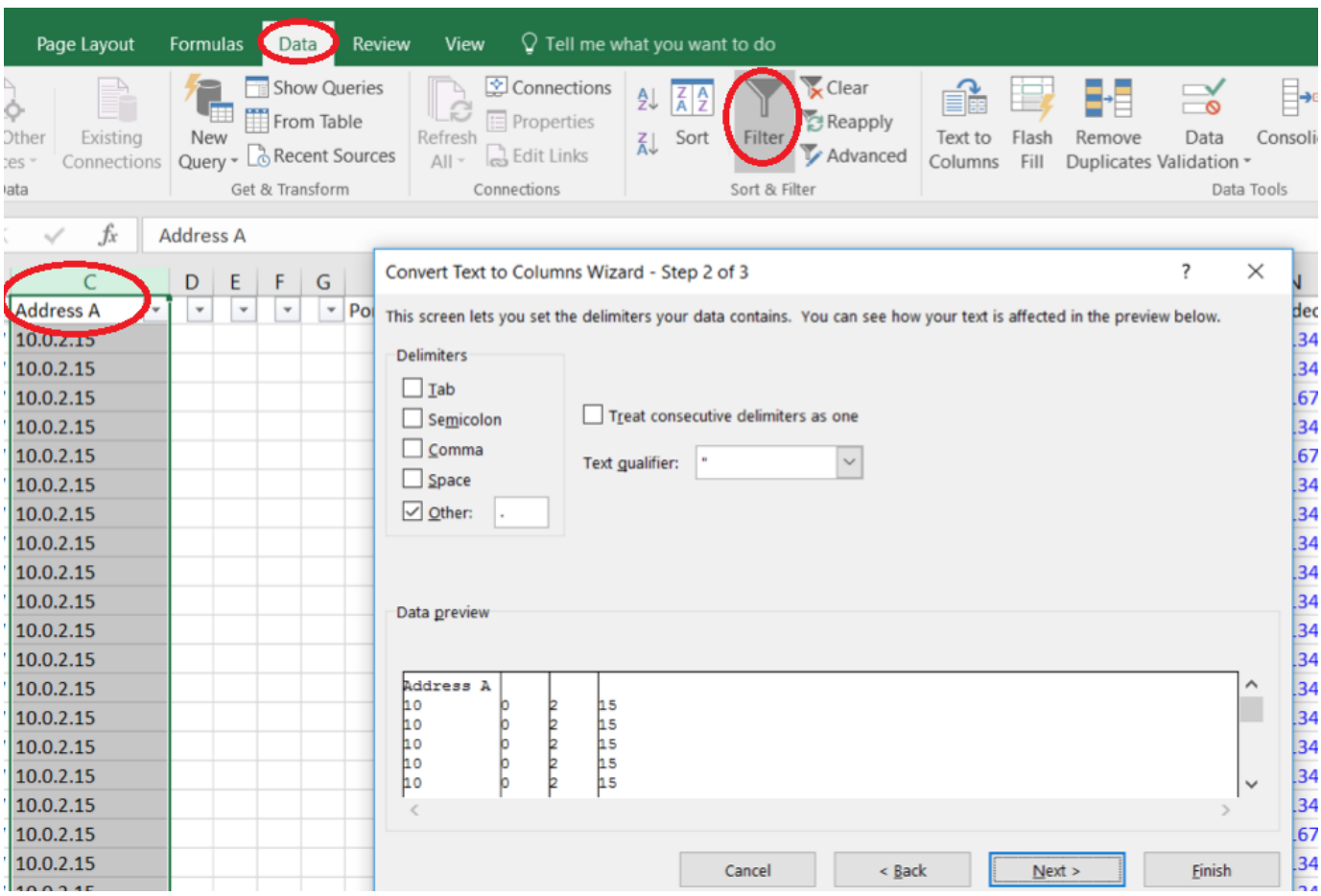
Records removed:

1- All address A values with 0.0.0.0
2- Port B Values with 53 (DNS)

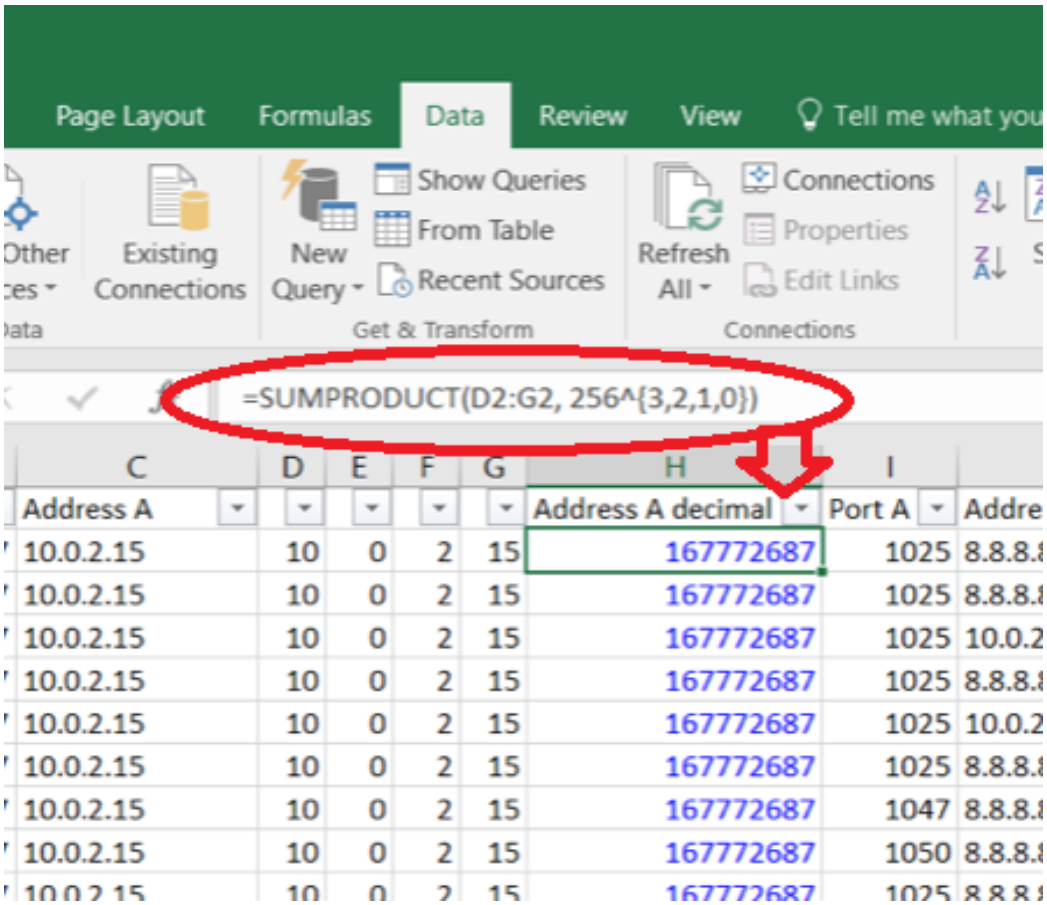
Columns removed (best results):

1- Packets
2- Bytes
3- Duration
4- Rel. start

Address A and **Address B** column IP addresses were split into 4 new columns using:
Excel / Data / Text to Columns feature with a "." delimiter.



The IP address was then converted to decimal using the following equation, followed by deletion of the extra redundant columns:
=SUMPRODUCT(D2:G2, 256^{3,2,1,0})



Second:

The data was split into **training** and **test** datasets that contained conversation data for an equal number of source goodware and malware PCAP files. The actual number of records differed between the datasets due to the difference in conversation data generated.

Third:

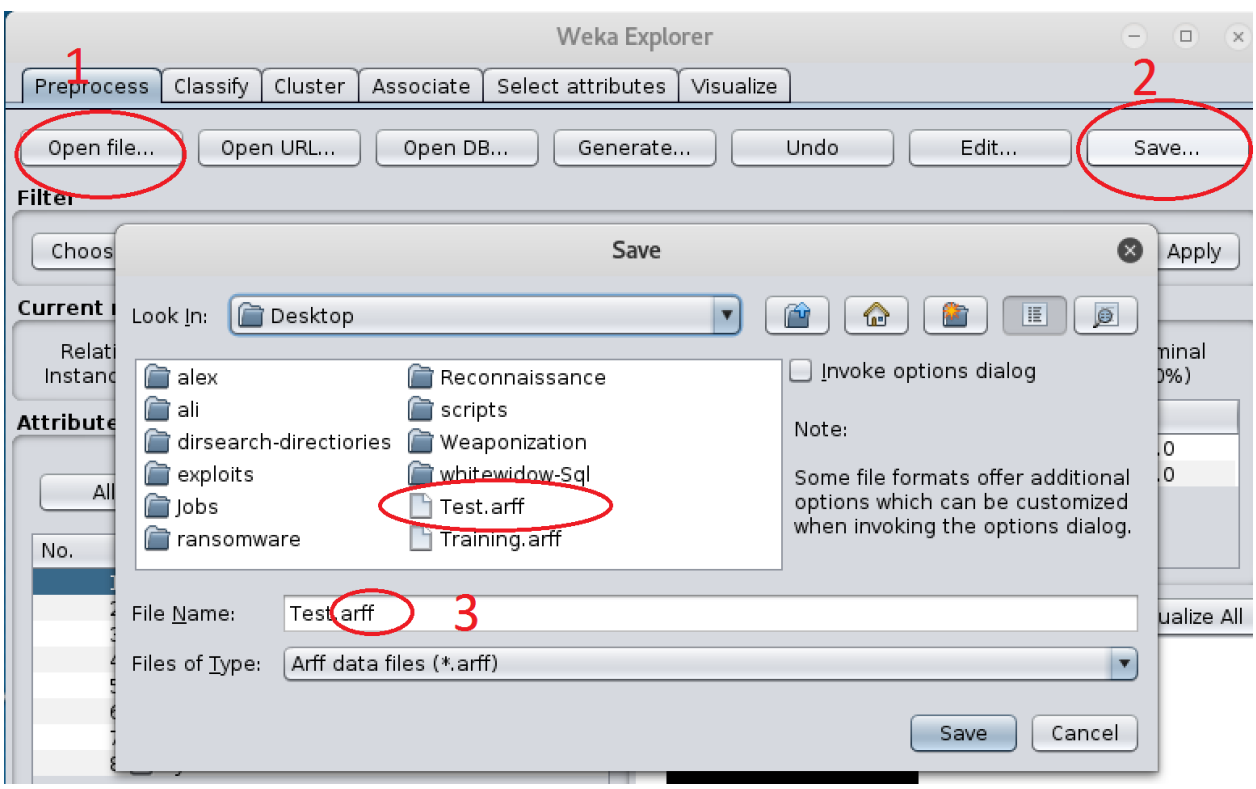
The hash column was removed to give us final column headings:

1- Label
2- Protocol
3- Address A
4- Port A
5- Address B
6- Port B
7- Packets A - B
8- Bytes A - B
9- Packets B - A
10- Bytes B – A

	A	B	C	D	E	F	G	H	I	J
1	Label	Protocol	Address A	Port A	Address B	Port B	Packets A - B	Bytes A - B	Packets B - A	Bytes B - A

Fourth:

The csv files were opened in WEKA Explorer and saved as **arff** files for further processing within WEKA.



Fifth:

ROC curves below produced from different classifiers, which we used in our experiment to compare our classifier performances. The picture below shows the use of Knowledge-flow environment of Weka to generate multiple ROC curves for more than one classifier, which we used in our paper.