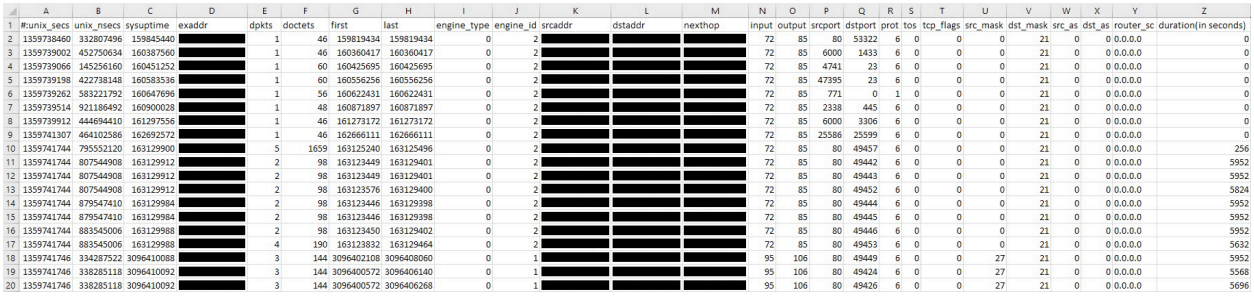
Profiling Internet Users

The source of data for this project is Cisco NetFlow version 5, which is one of the most popular technologies to collect IP traffic. Many parameters can be extracted from the source data including; Packets, Octets, beginning and ending of each flow, source and destination port numbers, source and destination IP addresses and many other variables which are included in the following figure.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	#unix_secs	unix_nsecs	sysuptime	exaddr	dpkts	doctets	first	last	engine_type	engine_id	srcaddr	dstaddr	nexthop	input	output	srcport	dstport	prot	tos	tcp_flags	src_mask	dst_mask	src_as	dst_as	router_sc	duration(in seconds)
2	1359738460	332807496	159845440	[redacted]	1	46	159819434	159819434	0	2	[redacted]	[redacted]	[redacted]	72	85	80	53322	6	0	0	0	21	0	0	0.0.0.0	0
3	1359739002	452750634	160387560	[redacted]	1	46	160360417	160360417	0	2	[redacted]	[redacted]	[redacted]	72	85	6000	1433	6	0	0	0	21	0	0	0.0.0.0	0
4	1359739066	145256160	160451252	[redacted]	1	60	160425695	160425695	0	2	[redacted]	[redacted]	[redacted]	72	85	4741	23	6	0	0	0	21	0	0	0.0.0.0	0
5	1359739198	422738148	160583536	[redacted]	1	60	160556256	160556256	0	2	[redacted]	[redacted]	[redacted]	72	85	47395	23	6	0	0	0	21	0	0	0.0.0.0	0
6	1359739262	583221792	160647696	[redacted]	1	56	160622431	160622431	0	2	[redacted]	[redacted]	[redacted]	72	85	771	0	1	0	0	0	21	0	0	0.0.0.0	0
7	1359739514	921186492	160900028	[redacted]	1	48	160871897	160871897	0	2	[redacted]	[redacted]	[redacted]	72	85	2338	445	6	0	0	0	21	0	0	0.0.0.0	0
8	1359739912	444694410	161297556	[redacted]	1	46	161273172	161273172	0	2	[redacted]	[redacted]	[redacted]	72	85	6000	3306	6	0	0	0	21	0	0	0.0.0.0	0
9	1359741307	464102586	162692572	[redacted]	1	46	162666111	162666111	0	2	[redacted]	[redacted]	[redacted]	72	85	25586	25599	6	0	0	0	21	0	0	0.0.0.0	0
10	1359741744	795552120	163129900	[redacted]	5	1659	163125240	163125496	0	2	[redacted]	[redacted]	[redacted]	72	85	80	49457	6	0	0	0	21	0	0	0.0.0.0	256
11	1359741744	807544908	163129912	[redacted]	2	98	163123449	163129401	0	2	[redacted]	[redacted]	[redacted]	72	85	80	49442	6	0	0	0	21	0	0	0.0.0.0	5952
12	1359741744	807544908	163129912	[redacted]	2	98	163123449	163129401	0	2	[redacted]	[redacted]	[redacted]	72	85	80	49443	6	0	0	0	21	0	0	0.0.0.0	5952
13	1359741744	807544908	163129912	[redacted]	2	98	163123576	163129400	0	2	[redacted]	[redacted]	[redacted]	72	85	80	49452	6	0	0	0	21	0	0	0.0.0.0	5824
14	1359741744	879547410	163129984	[redacted]	2	98	163123446	163129398	0	2	[redacted]	[redacted]	[redacted]	72	85	80	49444	6	0	0	0	21	0	0	0.0.0.0	5952
15	1359741744	879547410	163129984	[redacted]	2	98	163123446	163129398	0	2	[redacted]	[redacted]	[redacted]	72	85	80	49445	6	0	0	0	21	0	0	0.0.0.0	5952
16	1359741744	883545006	163129988	[redacted]	2	98	163123450	163129402	0	2	[redacted]	[redacted]	[redacted]	72	85	80	49446	6	0	0	0	21	0	0	0.0.0.0	5952
17	1359741744	883545006	163129988	[redacted]	4	190	163123832	163129464	0	2	[redacted]	[redacted]	[redacted]	72	85	80	49453	6	0	0	0	21	0	0	0.0.0.0	5632
18	1359741746	334287522	3096410088	[redacted]	3	144	3096402108	3096408060	0	1	[redacted]	[redacted]	[redacted]	95	106	80	49449	6	0	0	27	21	0	0	0.0.0.0	5952
19	1359741746	338285118	3096410092	[redacted]	3	144	3096400572	3096406140	0	1	[redacted]	[redacted]	[redacted]	95	106	80	49424	6	0	0	27	21	0	0	0.0.0.0	5568
20	1359741746	338285118	3096410092	[redacted]	3	144	3096400572	3096406268	0	1	[redacted]	[redacted]	[redacted]	95	106	80	49426	6	0	0	27	21	0	0	0.0.0.0	5696

To preserve the privacy, all the IP addresses are removed from the data. 54 Excel files are included in the project which each file corresponds to one user. Data is captured a month long period. In average, the number of flows for each subject over a week worth of data is more than 7000.

You may download the files from the link below:

https://drive.google.com/drive/folders/0Bw41Rn20xkcRenJmMEhtbzhVX0k?usp=sharing

In this project, we want to demonstrate if the Internet usage of each subject is statistically indistinguishable when compared to the Internet usage of the same subject over time, while simultaneously being statistically distinguishable when compared to Internet usage of other subjects. Subsequently, we want to study how the time window chosen for profiling affects the answers to the above problem. You can implement a profile for each user based on many criteria; however, it is highly suggested to use the ratio of octets/duration as the sole parameter for profiling. This parameter will henceforth be referred to as Internet Duration in this document

Write a program in any language that you are familiar in order to get network data as the input and do the statistical analysis to find the distinguishability or indistinguishability. Each file should be opened and compared with rest of the files. Good to mention that each file should be split in chunks of data. For example, you can split in chunks of 24 hours or as low as 15 seconds. You should find the shortest window that each user is statistically indistinguishable when compared to the Internet usage of the same subject, while simultaneously being statistically distinguishable when compared to Internet usage of other subjects.

For the statistical analysis part, you may use the flowing steps.

1. First step is to find the human readable date and time based on the columns in the NetFlow data. Below is a screenshot from one of the files included in the project. As it is clear, there are two columns named “Real First Packet” and “Real End Packet” that are date and time in epoch. Based

on these two columns, human readable date and time can be calculated. For learning about epoch time and how to convert it to human readable date and time look at the link below.

https://www.epochconverter.com/

At the link, you may find the way in different programing languages to convert epoch to human readable date and time. For example if you are using C#, you may use:

```
private string epoch2string(int epoch)

{

return new DateTime(1970, 1, 1, 0,
0,0,DateTimeKind.Utc).AddSeconds(epoch).ToShortDateString();

}
```

An example of conversion:

Real First Packet: 1359698407671 Human readable date and time: Friday, February 1, 2013 6:00:07.671 AM

unix_secs	sysuptime	dpkts	doctets	doctets/ dpkts	Real First Packet	Real End Packet	first	last	Duration
1359698757	120142852	1	46	46	1359698732619	1359698732619	120118471	120118471	0
1359700084	121469536	1	46	46	1359700059060	1359700059060	121444596	121444596	0
1359701391	122776296	1	46	46	1359701365073	1359701365073	122750369	122750369	0
1359701391	122776336	1	46	46	1359701365095	1359701365095	122750431	122750431	0
1359701391	122776372	1	46	46	1359701365060	1359701365060	122750432	122750432	0
1359701391	122776384	1	46	46	1359701365048	1359701365048	122750432	122750432	0
1359701391	122776424	1	46	46	1359701365070	1359701365070	122750494	122750494	0
1359701391	122777076	1	46	46	1359701364417	1359701364417	122750493	122750493	0
1359703224	124610212	1	60	60	1359703199776	1359703199776	124585988	124585988	0
1359703244	124630168	6	276	46	1359703218124	1359703227532	124604292	124613700	9408
1359703689	125074748	1	48	48	1359703659241	1359703659241	125044989	125044989	0

In this project, you may use only weekdays. You don't need to consider weekeneds.

2. After splitting the data into chunks, you need to find the correlation between them. Since you are comparing correlations across weeks, you should split the months' worth of Internet usage data into four chunks each for four weeks for all subjects. A brief snapshot of two weeks data for two subjects across time is shown in the following figure. In the following sample, window of 227 seconds is chosen. Column on the left is showing for a sample *"User A"* and column on the right is showing a sample data for *"User B"*. Each row is representing a window. For an instance, first row is representing data for Monday from 00:00:00am to 00:03:47am which is a 227-second window with the value of 6.3972 for the parameter of octets/duration. Similar procedure was done until Friday 11:56:13pm to 00:00:00am.

	User A		User B		
	Time	Octets/Duration	Time	Octets/Duration	
Week 1	Monday (00:00:00am-00:03:47am)	6.3972	Monday (00:00:00am-00:03:47am)	0.0302	Week 1
	⋮	⋮	⋮	⋮	
	Monday (11:56:13pm-00:00:00am)	4.9369	Monday (11:56:13pm-00:00:00am)	13.7590	
	Tuesday (00:00:00am-00:03:47am)	5.0646	Tuesday (00:00:00am-00:03:47am)	1.4598	
	⋮	⋮	⋮	⋮	
	Tuesday (11:56:13pm-00:00:00am)	4.2846	Tuesday (11:56:13pm-00:00:00am)	0.7783	
	Wednesday (00:00:00am-00:03:47am)	5.7988	Wednesday (00:00:00am-00:03:47am)	2.6305	
	⋮	⋮	⋮	⋮	
	Wednesday (11:56:13pm-00:00:00am)	2.3436	Wednesday (11:56:13pm-00:00:00am)	6.2205	
	Thursday (00:00:00am-00:03:47am)	2.4772	Thursday (00:00:00am-00:03:47am)	0.0000	
	⋮	⋮	⋮	⋮	
	Thursday (11:56:13pm-00:00:00am)	3.1775	Thursday (11:56:13pm-00:00:00am)	0.0000	
	Friday (00:00:00am-00:03:47am)	4.8082	Friday (00:00:00am-00:03:47am)	9.1049	
	⋮	⋮	⋮	⋮	
	Friday (11:56:13pm-00:00:00am)	5.0530	Friday (11:56:13pm-00:00:00am)	0.0000	
Week 2	Monday (00:00:00am-00:03:47am)	6.4694	Monday (00:00:00am-00:03:47am)	2.0793	Week 2
	⋮	⋮	⋮	⋮	
	Monday (11:56:13pm-00:00:00am)	4.3542	Monday (11:56:13pm-00:00:00am)	36.1807	
	Tuesday (00:00:00am-00:03:47am)	8.2608	Tuesday (00:00:00am-00:03:47am)	4.2334	
	⋮	⋮	⋮	⋮	
	Tuesday (11:56:13pm-00:00:00am)	8.1370	Tuesday (11:56:13pm-00:00:00am)	4.3147	
	Wednesday (00:00:00am-00:03:47am)	12.6390	Wednesday (00:00:00am-00:03:47am)	4.8411	
	⋮	⋮	⋮	⋮	
	Wednesday (11:56:13pm-00:00:00am)	12.6685	Wednesday (11:56:13pm-00:00:00am)	3.4661	
	Thursday (00:00:00am-00:03:47am)	11.6330	Thursday (00:00:00am-00:03:47am)	14.3444	
	⋮	⋮	⋮	⋮	
	Thursday (11:56:13pm-00:00:00am)	14.2283	Thursday (11:56:13pm-00:00:00am)	1.1753	
	Friday (00:00:00am-00:03:47am)	13.3379	Friday (00:00:00am-00:03:47am)	7.6747	
	⋮	⋮	⋮	⋮	
	Friday (11:56:13pm-00:00:00am)	17.3506	Friday (11:56:13pm-00:00:00am)	10.0920	

Note that some flows have a duration of 0 which due to the reason that the granularity is too short, the duration is 0 millisecond. Since you need to divide octets over duration and dividing by zero is undefined, you may not consider flows with 0 duration.

3. At this step you need to calculate the correlation coefficient values. There are three main type of correlation coefficient that for this project it is recommended to use the Spearman's correlation coefficient. You need to find three correlation values of r_{1a2a}, r_{1a2b} and r_{2a2b}. Numbers are showing the weeks and characters are showing the subjects. For example r_{1a2a} denotes the Spearman's correlation coefficient between Internet usage of *"Subject a"* for week 1 with Internet usage of *"Subject a"* for week 2. Similarly, r_{1a2b} denotes the Spearman's correlation coefficient between Internet usage of *"Subject a"* for week 1 with Internet usage of *"Subject b"* for week 2 and r_{2a2b} denotes the Spearman's correlation coefficient between Internet usage of *"Subject a"* for week 2 with Internet usage of *"Subject b"* for week 2.

4. Based on the correlation values which are calculated in the previous step, the main part of this project can be done. For the statistical framework of this project, Meng, Rosenthal, and Rubins Z Test Statistic (MRR-Z test) can be employed to find the value of Z. Required formulas are included below.

$$Z = [Z_{1a2a} - Z_{1a2b}] * \frac{\sqrt{[N-3]}}{2 * [1 - r_{2a2b}] * h}$$

$$Z_{1a2a} = \frac{1}{2} log \frac{1 + r_{1a2a}}{1 - r_{1a2a}}$$

$$Z_{1a2b} = \frac{1}{2} log \frac{1 + r_{1a2b}}{1 - r_{1a2b}}$$

$$h = \frac{1 - [f * rm^2]}{1 - rm^2}$$

$$f = \frac{1 - r_{2a2b}}{2 * [1 - rm^2]}$$

$$rm^2 = \frac{r_{1a2a}^2 + r_{1a2b}^2}{2}$$

N: Sample size of the data set

5. Based on the Z value calculated from the previous part, the corresponding P-value can be computed as follows:

$$P = 1 - \Phi(Z)$$

where $\Phi(Z)$ is the cumulative distribution function of standard normal distribution. You may use the following function in your code to find the P-value.

```
static double PFunction(double z)
{
    double p = 0.3275911;
    double a1 = 0.254829592;
    double a2 = -0.284496736;
    double a3 = 1.421413741;
    double a4 = -1.453152027;
    double a5 = 1.061405429;

    int sign;
    if (z < 0.0)
        sign = -1;
    else
        sign = 1;

    double x = Math.Abs(z) / Math.Sqrt(2.0);
    double t = 1.0 / (1.0 + p * x);
    double erf = 1.0 - (((((a5 * t + a4) * t) + a3)
      * t + a2) * t + a1) * t * Math.Exp(-x * x);
    return 0.5 * (1.0 + sign * erf);
}
```

6. Finally, based on the value that calculated from the previous step, you can decide that two users are distinguishable or indistinguishable from each other. When $P \leq 0.05$ means that correlation coefficient calculated for Internet usage patterns for an unknown subject (say b) is significantly smaller than that for a known subject (say a) and as such *"subject b"* will be identified as a subject distinct from *"subject a"*. On the contrary, when $P > 0.05$, indicates that correlation coefficient calculated for Internet usage patterns for an unknown subject (say b) is not significantly smaller than that for a known subject (say a), and as such *"subject b"* will be identified as indistinguishable from *"subject a"*. To finish, you need to write a report and briefly explain the procedure and include a table, which shows different time windows that used in the project and average number of matches for each window.