On the Feasibility of Profiling Internet Users based on Volume and Time of Usage

Soheil Sarmadi*, Mingyang Li† and Sriram Chellappan*
*Department of Computer Science and Engineering, University of South Florida, Tampa, FL, USA
sarmadi@mail.usf.edu, sriramc@usf.edu
†Department of Industrial and Management Systems Engineering, University of South Florida, Tampa, FL, USA
mingyangli@usf.edu

***Abstract*—In this paper, we address the issue of profiling users over the Internet using meta-data logs derived from network flow data (hence preserving a high degree of privacy). In this broader context, we specifically aim to empirically demonstrate that Internet volume and time of usage of humans do exhibit repeatable behavior over time. In our experimental study, Internet usage statistics of octets and duration (collected via privacy-preserving NetFlow records) of 66 student subjects in a college campus was recorded for a month. Subsequently, using state-of-the-art statistical techniques, we demonstrate how the Internet usage of any particular subject is highly-correlated with usage of the same subject over multiple time scales, while simultaneously being distinct from usage of other subjects. We derive interesting and practically useful trends on the relationship between the degree of distinguishability and the window time chosen to do the profiling. We also present discussions on the practical applications of this new study.**

***Index Terms*—Profiling, Network Forensics, Network Management, Privacy**

I. Introduction

Understanding user behavior profiles over the Internet is a topic of serious interest today. Such profiles can be individual or group based. For instance, for optimal deployment and management of network resources, service providers routinely profile network traffic of user groups to derive trends [1] [2]. Additionally, there is an ardent interest in the cyber crime community now to understand how criminals in cyber space use the Internet [3] [4]. In the realm of network security, and specifically authentication, there are efforts to model prior profiles of Internet users, and later use these derived models for authentication using challenge-response mechanisms [5]. Finally, in the broad field of cyber-psychology, there is a lot of interest now to associate Internet usage profiles with psychological disorders like depression, anxiety etc. [6]–[8].

In this paper, we make new contributions to behavioral based profiling of Internet users, in a manner that preserves a high degree of privacy. Specifically, using just Internet usage times and octets, and applying strong statistical techniques, we demonstrate how Internet usage (specifically octets/duration) of a subject does show a high degree of self-similarity for the same subject, while being distinct across subjects over varying time scales. Specifically, our contributions are:

a. Real Internet Usage Data Collected via NetFlow Logs: In the entire month of February, we collected Campus NetFlow logs of 66 undergraduate (UG) students in a college campus, all whose ids were anonymized. We point out that Internet usage logs of campuses, and most organizations across the globe, are routinely being collected for monitoring and troubleshooting purposes. Specifically the Internet usage logs collected as part of this study were NetFlow logs that provide us information on Internet flows for each subject, from which usage times, octets, packets, port numbers and protocols can be gleaned. It is important to see that NetFlow logs are highly-privacy preserving since content of Internet usage is never logged (e.g., contents of emails, or chats, or file downloads are never logged), but only statistics are logged. Subsequently, we preprocessed the logs to identify times of Internet usage in seconds, and the volume in bytes (denoted as octets) for the entire month.

b. Statistical Analysis of Temporal Internet Usage: Subsequently, we conducted a statistical analysis on the month's worth of Internet usage data (specifically, octets/duration) of each subject to answer the following questions. First, we wanted to see if each subject's usage data for days in one week exhibits statistically-strong correlations with the same subject's usage data for the same day over multiple weeks. Second, we wanted to see if each subject's usage data for days in one week is statistically different from that of other subjects' for the same day over multiple weeks. Finally, we wanted to see the impact of how the above correlations are affected based on changing the time window chosen to develop profiles. To answer these questions, we employ the classical Meng, Rosenthal, and Rubins Z Test Statistic (MRR-Z), which as we argue later is a widely used test and highly relevant to our problem scope.

c. Our Results: Our detailed statistical analysis reveals interesting and practically useful insights. First, we find that across multiple time windows for any weekday (i.e., 24-hour, 20-hour, 16-hour, 12-hour, 10-hour, 6-hour, 3-hour, 1-hour, 30-minute, 15-minute, 5-minute, 227-second, 30-second, and 15-second), each subject's Internet usage (i.e., octets/duration) is strongly correlated with the same day's usage for the same subject across all weeks. Interestingly, we also find that when the time windows to profile are longer, Internet usage of more subjects statistically correlate with those of any given subject. Also, if the profiling time goes down, there is a decreasing trend in the number of other subjects whose Internet usage correlates with those of any given subject, up to a point after

978-1-5386-2098-4/17/$31.00

which the number of subjects whose Internet usage correlate with a given subject starts to go up. Plotting the number of subjects whose Internet usage match those of any given subject versus the time window to profile yields a **U**-shaped curve (with the minimum point being a profiling window of 227 seconds in our study. Leveraring from these insights, we also present practical impacts of our work at the end).

II. Related Work

We present a brief overview of important related work. Due to space limitations, a comprehensive survey is not presented.

In [1] and [2], network profiling for anomalies detection are proposed. While the work in [1] provides a practical tutorial for profiling, the work in [2] identifies various granularities of destination network, host-pair, or host and port quadruple as markers for profiling Internet traffic. In other work in [3] and [4], Internet usage profiling was studied from the cybercrime point of view, wherein new methods are proposed to profile cyber criminals towards aiding subsequent network forensics.

In [5], a scheme called ActivPass is proposed [5] where the idea is to extract passwords from a user's daily activity logs, such as her Facebook activity, phone call activity etc. some of which are memorable, but unpredictable to others. Using challenge response mechanisms based on prior derived profiles, users are authenticated. In [6]–[8], work is presented that addresses a problem of urgent interest, namely understanding the relationship between mental health (like depression, stress, anxiety etc.) and Internet usage. Specifically, using statistical and machine learning techniques, this related work aims to derive models for correlating Internet usage times, Internet usage applications, social media posts, pictures on Instagram and more with various symptoms of mental health disorders.

To summarize, our work in this paper adds to this emerging field of behavioral-based Internet profiling. The problem we address, namely demonstrating the uniqueness of Internet usage times and octets of humans has not addressed before and one which has important practical applications.

III. Data Collection

In this section, we discuss the data collection aspect of our experimental study [1]. The source of data in this paper was Cisco NetFlow which is one of the most popular technologies to collect IP traffic. The data was collected from a sample of 66 UG students in a college campus network (with all identities anonymized) for the entire month of February. Briefly, NetFlow data collected from the campus network consists of several flows. In our study, NetFlow V5 was used, which contains numerous fields identified and described in Table I.

In order to distinguish each subject's data, the flows for each subject were identified based on the source IP address field and the same process continued for the entire month of data collected. The campus network where we collected data uses DHCP (Dynamic Host Configuration Protocol) provided IP addresses. As such, the IP address used by a subject at one time could be used by someone else later. Therefore, the process of extracting a subject's specific NetFlow logs begins by creating a mapping file and associating each subject with a set of assigned IP addresses, along with the start and end time stamps of each flow. This information is used by a backup daemon to extract subject-specific NetFlow information by filtering flows based on the source IP field. The mapping file is created by analyzing DHCP logs that include a subjects user-id, which is that subjects campus email address. Note that this process, summarized in Fig. 1, and was completely automated. Also, all ids were anonymized. A snapshot of NetFlow logs for a single subject is presented in Fig. 2, where each row denotes a single flow. We point out that on an average, the number of flows for each subject over a week worth of data was more than 7000. On an average, and each subject's Internet usage data via NetFlow logs for a week was around 3.75GB.

[1]The study was approved by the IRB at the participating campus.

TABLE I: Features collected via NetFlow logs

Feature	Description
unix_secs	Current count of seconds since 0000 UTC 1970
unix_nsecs	Residual nanoseconds since 0000 UTC 1970
sys_uptime	Current time in milliseconds since the export device booted
dPkts	Packets in the flow
dOctets	Total number of Layer 3 bytes in the packets of the flow
first	SysUptime at start of flow
last	SysUptime at the time the last packet of the flow was received
srcaddr	Source IP address
dstaddr	Destination IP address
srcport	TCP/UDP source port number or equivalent
dstport	TCP/UDP destination port number or equivalent
protocol	IP protocol bytes
src_mask	Source address prefix mask bits
dst_mask	Destination address prefix mask bits
src_as	Autonomous system number of the source, either origin or peer
dst_as	Autonomous system number of the destination, either origin or peer

Recall from Table I the fields that can be obtained from NetFlow logs. It is easy to see that some of these like Destination IP addresses, Ports and Protocols do provide information that is potentially useful for profiling. However, the focus of this paper is to demonstrate the feasibility of usage times and octets alone that preserve a high degree of privacy. As such, the NetFlow fields of interest to this study are:

- *duration:* This field is the amount of milliseconds from the start of flow to the end (converted to seconds herein).
- *octets:* This field is the number of Layer 3 bytes of the flow.

Note that in our study, octets and duration are integrated into a single parameter as a ratio (i.e., octets/duration) and denoted as Internet usage for the rest of the study. Usage profiles are also generated for this parameter only. Fig. 3 shows a snapshot of octets, duration and their ratio for a single subject, and similar tables are generated for all of the 66 subjects for the month in which NetFlow data was collected. The statistical analysis framework to generate profiles based on Internet usage is presented next.

IV. A Statistical Framework for Comparing Internet Usage across Subjects

In this section we present our overall statistical framework for profiling Internet subjects based the ratio of octets and

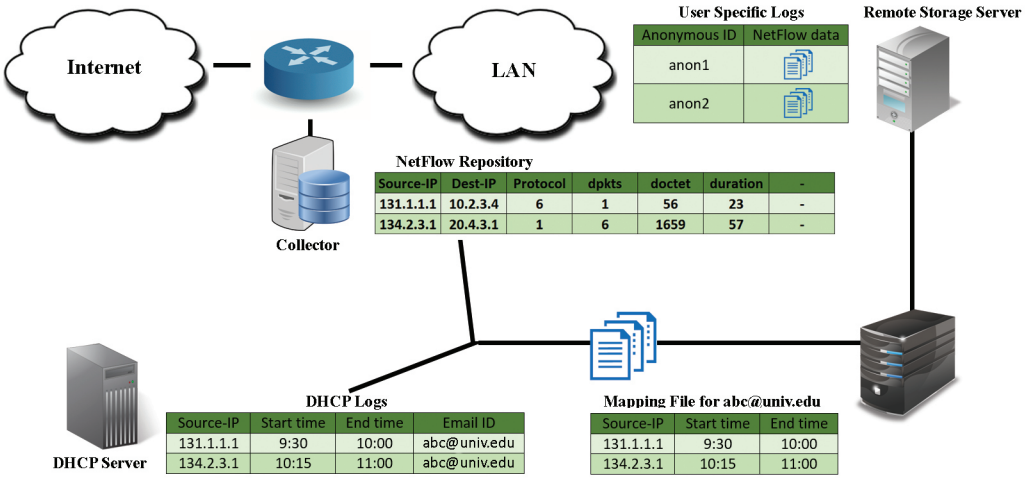

Fig. 1: Overall NetFlow data collection process

	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 ms)
2	1359702856	193227312	124241416	[redacted]	2	120	124215113	124218057	0	2	[redacted]	[redacted]	[redacted]	72	85	46442	25	6	0	0	0	21	0	0	0.0.0.0	2944
3	1359703688	840135996	125074060	[redacted]	1	48	125044157	125044157	0	2	[redacted]	[redacted]	[redacted]	72	85	33680	80	6	0	0	0	21	0	0	0.0.0.0	0
4	1359705840	898741974	3060504716	[redacted]	4	478	3060504462	3060504526	0	1	[redacted]	[redacted]	[redacted]	95	106	80	55309	6	0	0	27	21	0	0	0.0.0.0	64
5	1359705842	713967594	127227928	[redacted]	8	3305	127226652	127227100	0	2	[redacted]	[redacted]	[redacted]	72	85	443	55308	6	0	0	0	21	0	0	0.0.0.0	448
6	1359705840	990747714	3060504808	[redacted]	3	919	3060504463	3060504527	0	1	[redacted]	[redacted]	[redacted]	95	106	80	55310	6	0	0	27	21	0	0	0.0.0.0	64
7	1359705848	886749186	3060512704	[redacted]	4	971	3060511372	3060511820	0	1	[redacted]	[redacted]	[redacted]	95	106	80	55303	6	0	0	27	21	0	0	0.0.0.0	448
8	1359705849	190007874	127234404	[redacted]	10	2770	127233307	127233499	0	2	[redacted]	[redacted]	[redacted]	72	85	443	55302	6	0	0	0	21	0	0	0.0.0.0	192
9	1359705849	190007874	127234404	[redacted]	8	2634	127233563	127233691	0	2	[redacted]	[redacted]	[redacted]	72	85	443	55300	6	0	0	0	21	0	0	0.0.0.0	128
10	1359705849	190007874	127234404	[redacted]	8	2378	127233499	127233563	0	2	[redacted]	[redacted]	[redacted]	72	85	443	55301	6	0	0	0	21	0	0	0.0.0.0	64
11	1359705849	194005470	127234408	[redacted]	8	1890	127233690	127233754	0	2	[redacted]	[redacted]	[redacted]	72	85	443	55299	6	0	0	0	21	0	0	0.0.0.0	64
12	1359705849	194005470	127234408	[redacted]	7	2662	127233821	127233949	0	2	[redacted]	[redacted]	[redacted]	72	85	443	55297	6	0	0	0	21	0	0	0.0.0.0	128
13	1359705849	194005470	127234408	[redacted]	9	2430	127234077	127234269	0	2	[redacted]	[redacted]	[redacted]	72	85	443	55295	6	0	0	0	21	0	0	0.0.0.0	192
14	1359705849	194005470	127234408	[redacted]	8	1424	127231580	127231964	0	2	[redacted]	[redacted]	[redacted]	72	85	80	55304	6	0	0	0	21	0	0	0.0.0.0	384
15	1359705849	202000662	127234416	[redacted]	8	2698	127233757	127233821	0	2	[redacted]	[redacted]	[redacted]	72	85	443	55298	6	0	0	0	21	0	0	0.0.0.0	64
16	1359705849	202000662	127234416	[redacted]	10	2490	127233948	127234076	0	2	[redacted]	[redacted]	[redacted]	72	85	443	55296	6	0	0	0	21	0	0	0.0.0.0	128
17	1359705857	185995020	127242400	[redacted]	8	1890	127234394	127234522	0	2	[redacted]	[redacted]	[redacted]	72	85	443	55293	6	0	0	0	21	0	0	0.0.0.0	128
18	1359705857	185995020	127242400	[redacted]	10	4818	127236058	127236122	0	2	[redacted]	[redacted]	[redacted]	72	85	443	55279	6	0	0	0	21	0	0	0.0.0.0	64
19	1359705857	185995020	127242400	[redacted]	8	2394	127236122	127236186	0	2	[redacted]	[redacted]	[redacted]	72	85	443	55278	6	0	0	0	21	0	0	0.0.0.0	64
20	1359705857	190007874	127242404	[redacted]	8	1890	127234524	127234588	0	2	[redacted]	[redacted]	[redacted]	72	85	443	55292	6	0	0	0	21	0	0	0.0.0.0	64

Fig. 2: Snapshot of Real NetFlow logs for one subject (some entries are shaded intentionally)

	A	B	C	D	E	F	G
1	Start	Last	Start Time	Finish Time	Octets (bytes)	Duration (Seconds)	Octets/Duration
2	1360072262874	1360072954722	2/5/2013, 8:51:02 AM	2/5/2013, 9:02:34 AM	7151	692	10.33381503
3	1360072955042	1360073003502	2/5/2013, 9:02:35 AM	2/5/2013, 9:03:23 AM	525	48	10.9375
4	1360073034299	1360073192825	2/5/2013, 9:03:54 AM	2/5/2013, 9:06:32 AM	2392	158	15.13924051
5	1360073466733	1360073469244	2/5/2013, 9:11:06 AM	2/5/2013, 9:11:09 AM	21	3	7
6	1360073473445	1360073478053	2/5/2013, 9:11:13 AM	2/5/2013, 9:11:18 AM	92	5	18.4
7	1360073480267	1360073491426	2/5/2013, 9:11:20 AM	2/5/2013, 9:11:31 AM	46	11	4.181818182
8	1360073492979	1360073509619	2/5/2013, 9:11:32 AM	2/5/2013, 9:11:49 AM	75	17	4.411764706
9	1360073492434	1360073492434	2/5/2013, 9:11:32 AM	2/5/2013, 9:11:32 AM	5	0	0
10	1360073507927	1360073507927	2/5/2013, 9:11:47 AM	2/5/2013, 9:11:47 AM	0	0	0
11	1360073523469	1360073619983	2/5/2013, 9:12:03 AM	2/5/2013, 9:13:39 AM	333	96	3.46875

Fig. 3: Snapshot of Internet usage calculated for a single subject

duration. First, we present the formal problem statement, followed by the statistical framework.

A. *Our Problem Statement*

The broad problem statement addressed by this paper is the following. Given Internet usage of subjects (i.e., octets/ duration as presented in Fig. 3), 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 impacts the answers to the above problem.

Unfortunately, this objective not so straightforward to accomplish. Namely there are some challenges that are domain-specific which we need to be addressed. Recall that our subject population are college students, and the Internet usage data collected was over a campus network. As we know, college students have strict schedules each day of the week. For instance, while some students have classes Monday, Wednesday and Friday, other may have classes only on Tuesday and Thursday. A few others may have different class patterns. Naturally, the times and volume of Internet data usage across all weekdays will not be consistent for a single subject. Even within a single day, there is likely to be no Internet usage during class times for example. Furthermore, in weekends, different students may have different schedules, and some maybe be completely off campus at certain times (during which there will be no usage recorded on the campus networks). Therefore, in order to keep the ground truth data consistent, in this paper,

we focus only on data collected during weekdays for our subjects. We also attempted statistical comparisons across the same weekdays only for the data sets collected [2]. Note that there were no campus closings in the month in which we collected data.

B. Our Statistical Framework

1) Overview of Approach: To address the above problem, and overcome challenges, we employ a statistical analysis framework. For our data sets, since we are comparing correlations across weeks, we split the month's worth of Internet usage data into four chunks each for four weeks for all subjects for multiple time windows. A brief snapshot of two weeks data for two subjects across time is shown in Fig. 4, that presents Octets/Duration for every time window of 227 seconds (chosen as an example among multiple time windows). With this data, the Meng, Rosenthal, and Rubins Z Test Statistic (MRR-Z test) is leveraged for answering the above questions. This test is used to statistically evaluate and compare the significant difference of similarity measures among different subjects. Since the data used in this work is not normally distributed, the normality assumption widely-assumed in conventional statistical tests (such as Z-test and T-test) become invalid here [9]. Instead, MRR-Z test is employed in this paper due to its rigorous statistical property (e.g., asymptotic normality) and easy-to-compute form [10]. It has been applied in often areas, such as psychology and behavior science [11]–[13], to rigorously compare correlated correlation coefficients calculated from diverse sources of experimental data.

In the following discussions, without loss of generality, we present statistical analysis for comparing Internet usage data for two arbitrary subjects a and b (among the 66 subjects) over Weeks 1 and 2 only. The framework is the same when applied for all subjects across all weeks.

2) Methodology of the Meng, Rosenthal, and Rubins Z Test Statistic (MRR-Z): Recall that we want to determine the similarity of each subject's Internet usage over time, while also wanting to determine the corresponding dissimilarity with that of other subjects. As such, we propose to formulate the following null and alterative hypotheses as below. Essentially, the null hypothesis below makes the claim that for two Subjects a and b, across Weeks 1 and 2, they exhibit patterns of Internet usage that are statistically indistinguishable from each other. The alternative hypothesis makes the claims that the corresponding Internet usage patterns of two subjects are distinguishable from each other, as presented below.

$$H_0 : r_{1a2a} \leq r_{1a2b} \tag{1}$$

$$H_1 : r_{1a2a} > r_{1a2b} \tag{2}$$

[2]It is important to note that the time frame for profiling is based on the student population in this study, and does not take away the generality of our proposed techniques.

Note that in the above expressions, r_{1a2a} denotes the Spearman's rank correlation coefficient between Internet usages of Subject a for Week 1 with Internet usages of Subject a for Week 2; and r_{1a2b} denotes the Spearman's rank correlation between Internet usages of Subject a for Week 1 and Internet usages of Subject b for Week 2. Note that when we compare usage data of the same subject across weeks, then $a = b$. When we compare usage data for different subjects across weeks, then $a \neq b$. Also, the Spearman's rank correlation is used in this paper to derive correlations, due to the fact that our Internet usage data is not normally distributed. In addition, Spearman's correlation assesses monotonic relationships (whether linear or not) between two variables. When data are not bivariate normal, Spearman's correlation coefficient is often used as the index of correlation [14].

However, the statistical analysis we want to attempt has a challenge, since the correlation coefficients r_{1a2a} and r_{1a2b} cannot be directly determined in practice. However what we can obtain are the estimated correlation values $\widehat{r}_{1a2a}$ and $\widehat{r}_{1a2b}$ based on the sample data that we have. This is illustrated in Fig. 4, where the data for two different Subjects a and b for Weeks 1 and 2 are presented (where for this example, the time window chosen to compute octets/duration is every 227-seconds). In this example, the parameters $\widehat{r}_{1a2a}$ and $\widehat{r}_{1a2b}$ can be computed as the Spearman's correlation coefficient as

$$\widehat{r}_{xy} = 1 - \frac{6\Sigma d_i^2}{N(N^2-1)}, \tag{3}$$

where N denotes the number of Internet usage samples in the time slot for comparison (which is based on the window size chosen for profiling) and d_i denotes the difference between the ranks of corresponding values of usage for one subject and another subject. Note that the estimates for $\widehat{r}_{1a2a}$ and $\widehat{r}_{1a2b}$ are dependent since they are both computed based on the Internet usage of Subject a for Week 1. The MRR-Z test is specifically designed to compare such correlated coefficients with dependencies.

With these definitions, the MRR-Z statistical test for our hypothesis testing problem, can be done as follows by determining the parameter Z as:

$$Z = [Z_{1a2a} - Z_{1a2b}] * \frac{\sqrt{[N-3]}}{\sqrt{2 * [1 - \widehat{r}_{2a2b}] * h}}, \tag{4}$$

where $\widehat{r}_{2a2b}$ is the correlation coefficient between Week 2 of Subject a and Week 2 of Subject b and N is the sample size of the data set. Here Z_{1a2a} and Z_{1a2b} are Fisher's Z transformations of $\widehat{r}_{1a2a}$ and $\widehat{r}_{1a2b}$, which can be calculated respectively as:

$$Z_{1a2a} = \frac{1}{2} \log \frac{1 + \widehat{r}_{1a2a}}{1 - \widehat{r}_{1a2a}}, \tag{5}$$

$$Z_{1a2b} = \frac{1}{2} \log \frac{1 + \widehat{r}_{1a2b}}{1 - \widehat{r}_{1a2b}}, \tag{6}$$

The parameter h in Eq. 4 can be calculated based on the Eq. 7 with f and rm^2 computed in Eqs. 8 and 9, respectively

	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	

Fig. 4: Partitioning our data across weeks

$$h = \frac{1 - [f * rm^2]}{1 - rm^2}, \tag{7}$$

$$f = \frac{1 - \widehat{r}_{2a2b}}{2 * [1 - rm^2]}, \tag{8}$$

$$rm^2 = \frac{\widehat{r}^2_{1a2a} + \widehat{r}^2_{1a2b}}{2}. \tag{9}$$

The fisher transformation in Eqs. 5 and 6 helps transform sample correlation coefficients $\widehat{r}_{1a2a}$ and $\widehat{r}_{1a2b}$ closer to a normal distribution [15]. Under the null hypothesis, $Z_{1a2a} - Z_{1a2b}$ will further approximately follow normal distribution with mean 0 and standard deviation as:

$$Standard\ Deviation = 1/\frac{\sqrt{[N-3]}}{\sqrt{2 * [1 - \widehat{r}_{2a2b}] * h}}, \tag{10}$$

where standard deviation is calculated through h, f and rm^2 in Equations (7) to (9) [10].

Based on the MRR-Z test applied above to determine Z, the corresponding P-value can be computed as follows:

$$P = 1 - \Phi(Z), \tag{11}$$

where $\Phi(Z)$ is the cumulative distribution function of standard normal distribution i.e., $\Phi(Z) = P(Z \leq z)$.

Note here that based on a pre-specified significance level α (e.g, α=0.05), when $P \leq 0.05$, the null hypothesis of $\widehat{r}_{1a2a} \leq \widehat{r}_{1a2b}$ is rejected. This indicates 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 $P > 0.05$, the null hypothesis of $\widehat{r}_{1a2a} \leq \widehat{r}_{1a2b}$ cannot be rejected. It 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. Hence for our problem scope, the MRR-Z is applied across every pair of subjects for numerous time windows to determine the degree of distinguishability both within and across subjects, based on the computed P values.

Note that the significance level in this paper is set as a small value of 0.05. This setting can be interpreted as that when the null hypothesis is correct (i.e., $r_{1a2a} \leq r_{1a2b}$), the probability of making a mistake based on the MRR-Z test is smaller than 0.05. The significance level can be also adjusted to other smaller values, e.g., 0.01, based on requirements, although the statistical procedure remains the same.

V. Results of Statistical Analysis on Our Data Sets

We now present results of applying our statistical analysis framework for profiling based on Internet usage data sets. The time windows to compare correlation across subjects were chosen as 24-hour, 20-hour, 16-hour, 12-hour, 10-hour, 6-hour, 3-hour, 1-hour, 30-minute, 15-minute, 5-minute, 227-second, 30-second, and 15-second on all weekdays for the month in which Internet usage data was collected. Due to space limitations, presenting every possible result for all time windows across all four weeks is not possible. Instead, we present only a summary here, but the results are representative, and standard deviations from the average reported here were very low.

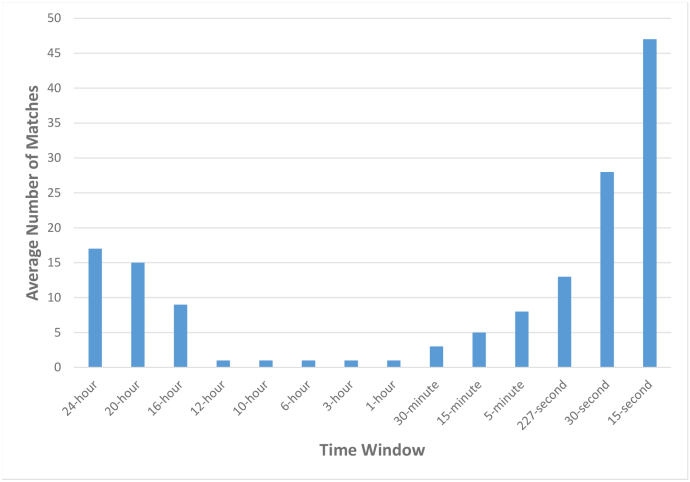

Fig. 5: Average No. of subject matches w.r.t. to different window sizes

Fig. 5 summarizes our results. The X-axis denotes different time windows chosen to compare correlations, while the Y-axis denotes the average number of subjects who were deemed to be statistically indistinguishable from every single subject using our statistical analysis above. For clarity of understanding, let us first present the manner in which we need to interpret our results. Namely, let us consider that subjects in our pool are labeled as $S_1, S_2, S_3, \ldots, S_{66}$ for 66 subjects. For each subject, we partition the Internet usage (octets/duration) across all time windows from 24 hours to 15 seconds for each weekday of the week for all four weeks. Note that a snapshot

of Internet usage for 227 seconds for a single subject was presented earlier in Fig. 4.

Consider for instance a time window $T = 5$ mins (or $T = 300$ seconds) for a single Subject S_1. We apply the MRR-Z test to determine the correlation between Internet usage data for every time window of 300 seconds for ever weekday in one week for Subject S_1 with the Internet usage of the corresponding time window of every other week for all users $S_1, S_2, S_3, \ldots, S_{66}$. We repeat this test across all subjects across all time windows for all weekdays and for all four weeks in which data was collected in order to determine the degree of similarity of usage within and across subjects. We also want to see how varying time windows chosen for correlation affect the similarities within and across users.

Fig. 5 presents our results. At the outset, we find that for all the time windows chosen, each subject's Internet usage data in any week demonstrated statistically significant correlations for the corresponding time window across all weekdays in all other weeks when compared with usage data of the same subject. This is important because this shows that even for very small time windows of 15 seconds, the Internet usage for every subject exhibits provably repeatable behavior. However this is only part of the puzzle, because we want to see how the statistical similarity of usage data when compared across subjects. As seen in in Fig. 5 for longer profiling time windows, a subject's usage data correlates with that of more subjects, and it slowly decreases as the prifiling window goes down to the point where only one subject is matched (which is always the same subject), and then the number of matched subjects starts to increase with further decrease in profiling time windows. These results in Fig. 5 are very insightful. They show that *only* with octets and durations, each subject's usage profile is unique when compared across one hour time windows. For much larger time windows, the granularity of octets and data is too large to characterize uniqueness. Similarly, when profiling time windows are too small, there is very limited usage data to derive characterize uniqueness for every subject.

Interestingly, for smaller time windows of 227 seconds, 5 minutes and 15 minutes, the number of matches is still very low. But in these time windows, those subjects whose usage profiles match those of other subjects are *only* for those cases where there is no usage data for those subjects (i.e., there are zero flows and octets for those subjects). When at-least one flow was present for a subject, the only statistically significant correlations obtained were for usage data of the same subject across weeks, while the usage data when compared for different subjects across weeks even in these small time windows did not correlate for any other week.

VI. Practical Applications and Conclusions

Demonstrating the feasibility of profiling users based on volume and time of usage alone, and deriving associated trends has not been attempted before. We present very briefly practical applications of our work. First, this work opens new possibilities of password-less authentication where usage volume and time at run-time can be compared with past usage to detect anomalies. Since we show profiling windows can be as small as 3 - 5 minutes, such a system will be practical. It is also possible now to build profiles based on roles in an organization - like security admin, database admin, network deployer etc., and use prior profiles for anomalies detection during run time when people abuse privileges. Personalized advertising and superior resource management for network deployers are also possible applications. However, for such applications to mature, we need significantly more data sets from many more users, with more diversity beyond campus settings, which is part of our current work. Specific tasks include deriving more privacy preserving features from traffic flow; looking into other tools that capture network traffic; enhancing subject diversity beyond campus settings; incorporating machine learning techniques for data processing and more.

References

[1] "Traffic profiling," http://www.cisco.com/c/en/us/td/docs/security/firepower/60/configuration/guide/fpmc-config-guide-v60/Creating˙Traffic˙Profiles.pdf.

[2] k. claffy, H. Braun, and G. Polyzos, "Internet traffic flow profiling," Applied Network Research, San Diego Supercomputer Center, Tech. Rep., Mar 1994.

[3] O. Wori, "Computer crimes: factors of cybercriminal activities," *International Journal of Advanced Computer Science and Information Technology*, vol. 3, no. 1, pp. pp–51, 2014.

[4] R. Saroha, "Profiling a cyber criminal," *International Journal of Information and Computation Technology*, vol. 4, no. 3, pp. 253–258, 2014.

[5] S. K. Dandapat, S. Pradhan, B. Mitra, R. Roy Choudhury, and N. Ganguly, "Activpass: your daily activity is your password," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015, pp. 2325–2334.

[6] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media." *ICWSM*, vol. 13, pp. 1–10, 2013.

[7] T. Martini, L. S. Czepielewski, A. Fijtman, L. Sodré, B. Wollenhaupt-Aguiar, C. S. Pereira, M. Vianna-Sulzbach, P. D. Goi, A. R. Rosa, F. Kapczinski *et al.*, "Bipolar disorder affects behavior and social skills on the internet," *PloS one*, vol. 8, no. 11, p. e79673, 2013.

[8] R. Katikalapudi, S. Chellappan, F. Montgomery, D. Wunsch, and K. Lutzen, "Associating internet usage with depressive behavior among college students," *IEEE Technology and Society Magazine*, vol. 31, no. 4, pp. 73–80, 2012.

[9] G. Casella and R. L. Berger, *Statistical inference*. Duxbury Pacific Grove, CA, 2002, vol. 2.

[10] X.-L. Meng, R. Rosenthal, and D. B. Rubin, "Comparing correlated correlation coefficients." *Psychological bulletin*, vol. 111, no. 1, p. 172, 1992.

[11] A. L. Duckworth and M. E. Seligman, "Self-discipline outdoes iq in predicting academic performance of adolescents," *Psychological science*, vol. 16, no. 12, pp. 939–944, 2005.

[12] P. Muris, H. Merckelbach, T. Ollendick, N. King, and N. Bogie, "Three traditional and three new childhood anxiety questionnaires: Their reliability and validity in a normal adolescent sample," *Behaviour research and therapy*, vol. 40, no. 7, pp. 753–772, 2002.

[13] J. M. Tybur, D. Lieberman, and V. Griskevicius, "Microbes, mating, and morality: individual differences in three functional domains of disgust." *Journal of personality and social psychology*, vol. 97, no. 1, p. 103, 2009.

[14] L. Myers and M. J. Sirois, "Spearman correlation coefficients, differences between," *Wiley StatsRef: Statistics Reference Online*, 2006.

[15] J. H. Steiger, "Tests for comparing elements of a correlation matrix," *Psychological bulletin*, vol. 87, no. 2, pp. 245–251, 1980.