2018 IEEE 32nd International Conference on Advanced Information Networking and Applications

A Statistical Framework to Forecast Duration and Volume of Internet Usage based on Pervasive Monitoring of NetFlow Logs

Soheil Sarmadi*, Mingyang Li†, 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 an important and practical problem - namely to forecast the duration and volume of Internet usage of a subject based on pervasive and unobtrusive of past history. Unfortunately though, profiling users can have privacy ramifications. In this paper, we present a statistical framework to forecast duration and volume of Internet usage of subjects via processing NetFlow logs from routers. Briefly, NetFlow logs are network level information of IP packets as they traverse a router, but they do not contain the packet payload. In our experimental study, Internet traffic logs of octets and durations of 66 subjects in a college campus were collected (via privacy-preserving NetFlow records) in a pervasive and unobtrusive manner for a month. By applying times series forecasting techniques, we demonstrate that predictions on duration and volume of usage at future times can be made based on past usage, with very good precision. Furthermore, our results also show that with more historical data, prediction accuracy improves further. We believe that our problem in this paper has not been addressed in the literature. We also believe that our contributions in this paper have important consequences in enabling privacy preserving techniques to manage network resources for administrators, cyber security via behavioral based authentication, and smarter advertising.**

***Index Terms*—Prediction, Privacy, Network Management, User behavior, Forecasting**

I. Introduction

Profiling user activity on the Internet is a topic that is important, with numerous scenarios and applications. Instances include a) deriving social media profiles on Facebook and Twitter for applications like friends matching, targeted advertising, demographic analysis [1]; b) deriving router profiles for network traffic management, resources deployment, attack/fault detection via anomalies [2]; c) deriving profiles for understanding apps usage on smart-phones [3]; d) deriving Internet usage profiles to predict mental health outcomes [4]–[7] and more.

In this paper, we make new contributions to user profiling on the Internet. The problem we address is both practical and important. Specifically, our problem is to determine the volume and duration of Internet usage of a subject based on knowledge of past historical data that was collected pervasively and unobtrusively. This problem has multiple uses in the realm of authentication (via anomaly detection), run-time management of network resources (via superior forecasting), and smarter advertising (again via forecasting usage times) and more. Unfortunately though, such a problem is hard to tackle mainly from a privacy perspective, since sharing of historical Internet data to derive profiles may be of concern to users.

To alleviate this problem, while still demonstrating the ability to forecast, we conduct an experimental study in this paper for forecasting Internet usage duration and volume from historical data collected via privacy preserving NetFlow logs from routers. Specifically, our contributions are:

a. Real Internet usage traffic collected via NetFlow: Internet usage of 66 undergraduate (UG) college students for the entire month of February was collected via NetFlow logs. All sensitive information was anonymized. Note that many organizations and corporations pervasively collect Internet traffic statistics for monitoring and troubleshooting purposes. To be more specific, the NetFlow logs used in this paper typically record information such as packets, octets, duration, port numbers and protocols of an IP packet that pass through a router. Note that the NetFlow logs are highly-privacy preserving since the content of Internet usage is never collected (e.g., content of emails, or chats, or file downloads are never logged). Only the statistics are collected. Subsequently, we processed the recorded data to identify the volume in bytes (denoted as octets) and duration of usage (as a notion of time) for the entire month for the purpose of this study.

b. Time Series Forecasting: Subsequently, we analyzed the entire month of Internet usage data to answer following questions. First, we wanted to see if each subject's usage data can be predicted based on the previous usage data. Second, we want to see if more statistics from previous usage data can help to predict future Internet usage more accurately. To answer these questions, we use Time Series Forecasting, which is one of the widely used methods of prediction in the literature.

c. Our Results: Our detailed prediction results reveal interesting and practical insights. First, we found that each week of Internet usage data can be predicted based on the previous weeks of usage data from the perspective of duration and volume. We also find that with more information to profile, accuracy of forecasting improves. Also, in general, profiling

1550-445X/18/$31.00 ©2018 IEEE
DOI 10.1109/AINA.2018.00077

and predicting future duration and volume of usage over 1-hour window performs best compared to other time windows.

The rest of the paper is organized as follows. In Section II, we discuss important related work. In Section III, we discuss the NetFlow Data collection process. Section IV presents in detail our statistical analysis for predicting future trend of Internet data. We present results of our analysis in Section V and practical applications are presented in Section VI. Finally, Section VII concludes the paper.

II. Related Work

In this section we present a brief overview of important related studies. Due to space limitations, a comprehensive survey is not presented.

Considering the importance of profiling Internet users, many innovations have been proposed in the literature. In [2], a parameterizable methodology for profiling Internet traffic flows at a variety of granularities is presented. Authors defined flows based on traffic satisfying various temporal and spatial locality conditions that were observed at internal points in a network. This methodology can solve some central problems with networking such as resource reservation at multiple service levels, usage based accounting, and the integration of IP traffic over an ATM fabric. Authors considered various granularities of a flow such as destination network, host-pair, or host and port quadruple in their analysis.

In [1], authors studied the feasibility of automatically extracting passwords from profiling daily activities like phone activity, Facebook activity etc. They have observed that infrequent activities can be memorable and also unpredictable. Authors launched an experiment with 70 subjects by including their Facebook activities, browsing history, call logs, texts. The proposed system could achieve a success rate of 95% to authenticate legitimate users and was compromised in 5.5% of cases. More recent related work in authentication is in [8], where the authors demonstrate that octets and duration of Internet usage tends to be repeatable over time for users, which can serve as new markers for authentication.

In [3], authors comprehensively studied the diverse usage patterns of smartphone applications via network measurements from a national level Tier-1 cellular network provider in the U.S. They observed that about 20% of the applications that are very popular are local because they are expected to serve local users such as news applications. They also found similarities across different applications in terms of geographic coverage, daily usage patterns, etc.

In [4], [5], [6], [7], the authors studied profiles of student Internet users with applications to mental healthcare. They demonstrate that Internet usage statistics like Octets, Usage Duration, Entropy of Usage, Chatting Usage etc. are significant different for students that suffer from mood problems, compared to those that dont. Furthermore, the privacy implications of such findings and their applications are also discussed.

In [9], authors studied Internet users behaviors toward Internet advertising and how it can be compared to advertising in general. They provided results which indicate that more respondents found Internet advertising to be informative and trustworthy than a demographically similar sample found general advertising.

To summarize, the work in this study adds to this emerging field of Internet usage profiling. The problem we address, namely forecasting the duration and volume of Internet usage from past profiles has not addressed before, and is one which has important practical applications.

III. Data Collection

In this section, we briefly introduce the data collection part of our project[1]. Data was collected using Cisco NetFlow which is one of the most well-known technologies to capture network traffic [10]. Data included a sample of 66 student subjects in a campus network (with all identities anonymized) for a month long period. In this experimental study, NetFlow version 5 was used that export many variables which are briefly described in Table I.

The collected NetFlow data has several flows which was categorized based on the source IP address field. Since the campus network which the data was collected uses DHCP (Dynamic Host Configuration Protocol) in order to assign IP address, each IP address can be used by different users at different times. As the result, a mapping file was established that contains a list of IP addresses which are assigned to each user. This file was created using DHCP logs that includes each subject's username, which is their email address. Mentioned information is used by a backup daemon to extract each subjects NetFlow data by filtering flows based on the source IP address variable. Fig. 1 presents the whole process that was done for collecting the NetFlow traffic. All ids and sensitive information were anonymized to preserve privacy. In average, a week of worth data contains more than 7000 flows and an Internet usage data of around 3.75GB.

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

As introduced in Table I, several fields can be captured from the data. Fig. 2 illustrates a snapshot of the real NetFlow traffic logs for one sample subject. However, in this paper we only considered octets and duration for forecasting. Forecasting other possible variables like port numbers, destination

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

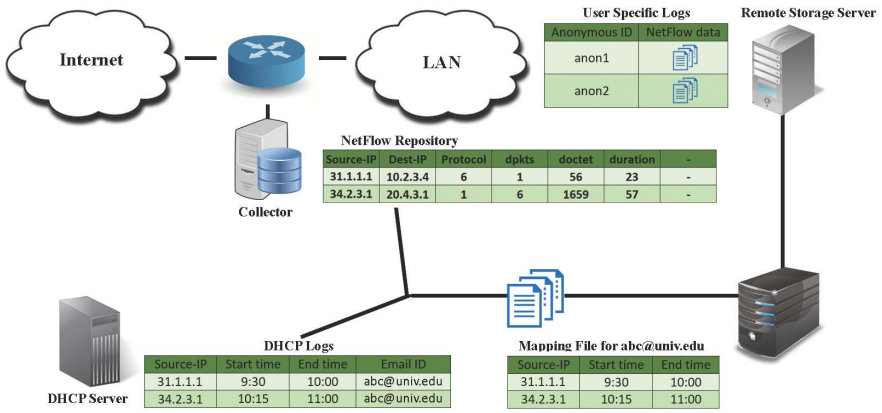

Fig. 1: Overall NetFlow data collection process

#: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
1361309853	869507646	1731233952		14	9030	1731207651	1731222307	0	2				72	85	80	62396	6	0	0	0	21	0	0	0.0.0.0	14656
1361309853	869507646	1731233952		11	768	1731207778	1731230050	0	2				72	85	45782	60067	6	0	0	0	21	0	0	0.0.0.0	22272
1361309853	989496558	1731234072		7	997	1731207075	1731211491	0	2				72	85	80	62374	6	0	0	0	21	0	0	0.0.0.0	4416
1361309853	989496558	1731234072		1	205	1731207074	1731207074	0	2				72	85	53	56104	17	0	0	0	21	0	0	0.0.0.0	0
1361309853	989496558	1731234072		4	2441	1731207074	1731207138	0	2				72	85	80	62376	6	0	0	0	21	0	0	0.0.0.0	64
1361309853	989496558	1731234072		9	9666	1731207137	1731207201	0	2				72	85	80	62382	6	0	0	0	21	0	0	0.0.0.0	64
1361309853	989496558	1731234072		9	9799	1731207137	1731207329	0	2				72	85	80	62384	6	0	0	0	21	0	0	0.0.0.0	192
1361309857	905501268	1731237988		5	1048	1731212386	1731232162	0	2				72	85	80	62406	6	0	0	0	21	0	0	0.0.0.0	19776
1361309857	905501268	1731237988		2	244	1731212193	1731212321	0	2				72	85	53	51234	17	0	0	0	21	0	0	0.0.0.0	128
1361309857	909498864	1731237992		1	123	1731212323	1731212323	0	2				72	85	53	63272	17	0	0	0	21	0	0	0.0.0.0	0
1361309869	897506076	1731249980		6	393	1731224930	1731247074	0	2				72	85	62134	60076	6	0	0	0	21	0	0	0.0.0.0	22144
1361309873	929502102	1731254012		9	618	1731225826	1731250722	0	2				72	85	36633	60066	6	0	0	0	21	0	0	0.0.0.0	24896
1361309874	13549104	1731254096		3	168	1731228770	1731238306	0	2				72	85	56210	60212	6	0	0	0	21	0	0	0.0.0.0	9536
1361309877	865510050	1731257948		10	693	1731231072	1731255392	0	2				72	85	55536	60069	6	0	0	0	21	0	0	0.0.0.0	24320
1361309877	989496558	1731258072		8	543	1731231010	1731246050	0	2				72	85	15704	60502	6	0	0	0	21	0	0	0.0.0.0	15040
1361309882	169531638	1731262252		4	243	1731237154	1731248162	0	2				72	85	45782	60067	6	0	0	0	21	0	0	0.0.0.0	11008
1361309914	274033680	1731294348		3	144	1731286562	1731287650	0	2				72	85	80	62419	6	0	0	0	21	0	0	0.0.0.0	1088
1361309914	274033680	1731294348		9	6135	1731286561	1731287905	0	2				72	85	80	62418	6	0	0	0	21	0	0	0.0.0.0	1344
1361309914	434029068	1731294508		4	2435	1731289695	1731289887	0	2				72	85	80	62422	6	0	0	0	21	0	0	0.0.0.0	192
1361309921	981989622	1731302056		8	543	1731277346	1731290210	0	2				72	85	56210	60212	6	0	0	0	21	0	0	0.0.0.0	12864
1361309922	106043100	1731302180		2	92	1731277345	1731277345	0	2				72	85	80	62410	6	0	0	0	21	0	0	0.0.0.0	0
1361309922	314024898	1731302388		5	691	1731295776	1731300896	0	2				72	85	80	62442	6	0	0	0	21	0	0	0.0.0.0	5120

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

addresses and more are possible, and is part of our future work, once privacy implications are better understood. Formally, the variables used in this research paper for forecasting are:

- *octets:* This parameter shows the number of Layer 3 bytes per flow.
- *duration:* This parameter shows the amount of milliseconds from the beginning of the flow to the ending (converted to seconds for ease of use).

The above procedure was completed for all the 66 student subjects in the study. Fig. 3 shows a snapshot for a single user from a new table that contains the processed Internet usage parameters in addition to converted date and time in epoch to human readable date and time for ease of use. At this step, different time windows have been employed to split the data into several parts. We have used windows in a range of 24 hours to 15 seconds for splitting our data. In prior research we have demonstrated that 1-hour window as preferable to partition the data and demonstrate self-similar behavior of Internet usage [8]. Interestingly, in this paper also we we found that 1-hour window can also predict the future trend more accurately than any other time window. We compare our forecasting results across multiple profiling windows later in the paper in the results section.

IV. Predicting Internet Users' traffic

In this section we present the framework to forecast the duration and volume of Internet usage based on past profiles of the same. Specifically, in this paper, we want to forecast the duration and volume of usage in each weekday of one week based on profiles derived for the same weekday in three prior weeks.

To design this framework, there are some challenges. For example, since our subjects are college students, they have a strict schedule. For an instance, one student can have classes on Mondays, Wednesdays and Fridays and others may have on Tuesdays and Thursdays. Also, there is a possibility that different weeks have different schedules. Furthermore, many students are off campus during the weekends so we only focused on data that captured during weekdays.

The main question that we want to answer in the area is the following. We want to see if any week of Internet usage can be predicted based on the previous weeks of Internet traffic. In addiction, we want to see if more previous data can help to predict more accurately. Finally, we want to find the best profiling window size to forecast.

To address the above problem, and overcome challenges, we use Time Series Forecasting technique [11]. This technique is considered to predict new data when the actual outcome may

Start	Last	Start Time	Finish Time	Octets (bytes)	Duration (Seconds)
1360072262874	1360072954722	2/5/2013, 8:51:02 AM	2/5/2013, 9:02:34 AM	7151	692
1360072955042	1360073003502	2/5/2013, 9:02:35 AM	2/5/2013, 9:03:23 AM	525	48
1360073034299	1360073192825	2/5/2013, 9:03:54 AM	2/5/2013, 9:06:32 AM	2392	158
1360073466733	1360073469244	2/5/2013, 9:11:06 AM	2/5/2013, 9:11:09 AM	21	3
1360073473445	1360073478053	2/5/2013, 9:11:13 AM	2/5/2013, 9:11:18 AM	92	5
1360073480267	1360073491426	2/5/2013, 9:11:20 AM	2/5/2013, 9:11:31 AM	46	11
1360073492979	1360073509619	2/5/2013, 9:11:32 AM	2/5/2013, 9:11:49 AM	75	17
1360073492434	1360073492434	2/5/2013, 9:11:32 AM	2/5/2013, 9:11:32 AM	5	0
1360073507927	1360073507927	2/5/2013, 9:11:47 AM	2/5/2013, 9:11:47 AM	0	0
1360073523469	1360073619983	2/5/2013, 9:12:03 AM	2/5/2013, 9:13:39 AM	333	96

Fig. 3: Snapshot of volume and duration for a single subject

not be known currently. Time Series Forecasting is employed in this paper due to its rigorous statistical property and easy-to-compute form. This modeling approach is particularly useful when there is little knowledge available on the underlying data generating process or there is no satisfactory explanatory model that relates the prediction variable to other explanatory variables. It has also been successfully applied in areas, such as statistics, pattern recognition, weather forecasting, earthquake prediction, and widely in any domain of applied science and engineering that involves temporal measurements. For our data sample, since we are predicting the fourth week based on previous three weeks, we split the month's worth of Internet traffic data into four parts for four weeks each for all 66 subjects as it can be seen in Fig. 4.

In the following discussions, without loss of generality, we present the technique for predicting Internet traffic data for an arbitrary subject (among the 66 subjects). The technique is the same when applied for all the subjects in the data sample. The process is summarized in algorithm 1. There are several variables which are briefly described in Table II

TABLE II: List of variables used in the Algorithm

Variable	Description
L:	denotes the amount of Internet traffic after splitting the data into windows
MA:	denotes the Moving Average
CMA:	denotes the Centered Moving Average
S:	denotes the seasonality component
DS:	denotes the deseasonalized variable
T:	denotes the trend component
Prediction:	denotes the predicted value for the Internet traffic

There are several different methods to do Time Series Forecasting. One of the traditional techniques that is employed in this paper is called Moving Average (MA) [12]. An observation in time series can be decomposed into three different components: the trend that shows long term direction, the seasonal that is systematic or calendar related movements and irregular that is unsystematic which is short term fluctuations related effect. This process is called decomposition model.

As mentioned earlier, one of the components is called seasonal effect that is a systematic and calendar related effect. Observed data needs to be seasonally adjusted since seasonal effects can hide the true underlying movement in the series. In addition, it can hide certain non-seasonal characteristics that may be of interest. The seasonal component consists of effects that are stable with respect to timing, direction and magnitude. Seasonality can be recognized by regularly spaced peaks that have a consistent direction and approximately the same magnitude in the time period. Another component of Time Series is called trend that is defined as the long term movement and is a reflection of the underlying level.

Algorithm 1 Internet traffic prediction

procedure PREDICTION(Data Flow)
  create windows of 1-hour for a specific day (e.g. Monday) across all three weeks: $L[1..72]$
  for $i = 13$ to 61 **do**
    $MA[i] = [L[i-12] + L[i-11] + L[i-10] + .. + L[i+11]]/24$
  end for
  for $j = 13$ to 60 **do**
    $CMA[j] = [MA[j] + MA[j+1]]/2$
  end for
  for $m = 1$ to 24 **do**
    $S[m] = [\frac{L[m]}{CMA[m]} + \frac{L[m+24]}{CMA[m+24]} + \frac{L[m+48]}{CMA[m+48]}]/3$
    $S[m + 24] = [\frac{L[m]}{CMA[m]} + \frac{L[m+24]}{CMA[m+24]} + \frac{L[m+48]}{CMA[m+48]}]/3$
    $S[m + 48] = [\frac{L[m]}{CMA[m]} + \frac{L[m+24]}{CMA[m+24]} + \frac{L[m+48]}{CMA[m+48]}]/3$
    $S[m + 72] = [\frac{L[m]}{CMA[m]} + \frac{L[m+24]}{CMA[m+24]} + \frac{L[m+48]}{CMA[m+48]}]/3$
  end for
  for $n = 1$ to 72 **do**
    $DS[n] = L[n]/S[n]$
  end for
  LinearRegression(DS[1 .. 72])
  for $o = 1$ to 96 **do**
    $T[o] = a + [b \times o]$
  end for
  for $q = 73$ to 96 **do**
    $Prediction = T[q] \times S[q]$
  end for
end procedure

Two structure are proposed for basic decomposition models; Additive and Multiplicative. Noteworthy, in this paper we used the classical time series multiplicative model as it is presented in Eq. (1).

$$x_t = Seasonal \times Trend \times Random \tag{1}$$

The term "Random" is often called "Irregular" in decompositions. Random or irregularity is canceled from the equation

by deseasonalising the process. Although, in this paper, to compute the prediction process, two components of seasonal and trend are employed as they are presented in Eq. (2). In short, trend is showing the long-run increase or decrease over time and seasonal is showing the short-term regular wave-like patterns. Different approaches are available, however in this paper we used a smoothing procedure called Moving Average (MA).

$$x_t = Seasonal \times Trend \tag{2}$$

User A

	Time	Octets (bytes)	Duration (seconds)	Time	Octets (bytes)	Duration (seconds)	
Week 1	Monday (00:00:00am-01:00:00am)	9738364	838	Monday (00:00:00am-01:00:00am)	11874	839	Week 3
	⋮	⋮	⋮	⋮	⋮	⋮	
	Monday (11:00:00pm-00:00:00am)	3498789	927	Monday (11:00:00pm-00:00:00am)	943731184	1264	
	Tuesday (00:00:00am-01:00:00am)	703	703	Tuesday (00:00:00am-01:00:00am)	7383	739	
	⋮	⋮	⋮	⋮	⋮	⋮	
	Tuesday (11:00:00pm-00:00:00am)	693	389	Tuesday (11:00:00pm-00:00:00am)	44527	847	
	Wednesday (00:00:00am-01:00:00am)	21976086	1257	Wednesday (00:00:00am-01:00:00am)	1662	404	
	⋮	⋮	⋮	⋮	⋮	⋮	
	Wednesday (11:00:00pm-00:00:00am)	85736474	1189	Wednesday (11:00:00pm-00:00:00am)	746603	754	
	Thursday (00:00:00am-01:00:00am)	2890	394	Thursday (00:00:00am-01:00:00am)	10231	327	
	⋮	⋮	⋮	⋮	⋮	⋮	
	Thursday (11:00:00pm-00:00:00am)	57991	458	Thursday (11:00:00pm-00:00:00am)	649325840	1843	
	Friday (00:00:00am-01:00:00am)	62415	859	Friday (00:00:00am-01:00:00am)	9340	205	
	⋮	⋮	⋮	⋮	⋮	⋮	
	Friday (11:00:00pm-00:00:00am)	24934	538	Friday (11:00:00pm-00:00:00am)	898765	957	
Week 2	Monday (00:00:00am-01:00:00am)	8614478	273	Monday (00:00:00am-01:00:00am)	2131	128	Week 4
	⋮	⋮	⋮	⋮	⋮	⋮	
	Monday (11:00:00pm-00:00:00am)	573768	1259	Monday (11:00:0)pm-00:00:00am)	119847	1192	
	Tuesday (00:00:00am-01:00:00am)	24793880	506	Tuesday (00:00:00am-01:00:00am)	137913581	927	
	⋮	⋮	⋮	⋮	⋮	⋮	
	Tuesday (11:00:00pm-00:00:00am)	7401160	845	Tuesday (11:00:0)pm-00:00:00am)	9332	307	
	Wednesday (00:00:00am-01:00:00am)	4268757	1267	Wednesday (00:0000am-01:00:00am)	193140805	1294	
	⋮	⋮	⋮	⋮	⋮	⋮	
	Wednesday (11:00:00pm-00:00:00am)	91094250	1273	Wednesday (11:0000pm-00:00:00am)	51771708	1071	
	Thursday (00:00:00am-01:00:00am)	29635143	137	Thursday (00:00:0)am-01:00:00am)	889	450	
	⋮	⋮	⋮	⋮	⋮	⋮	
	Thursday (11:00:00pm-00:00:00am)	318987	627	Thursday (11:00:00pm-00:00:00am)	763197621	1335	
	Friday (00:00:00am-01:00:00am)	45610264	385	Friday (00:00:00am-01:00:00am)	86649314	1240	
	⋮	⋮	⋮	⋮	⋮	⋮	
	Friday (11:00:00pm-00:00:00am)	13112567	2084	Friday (11:00:00pm-00:00:00am)	643179872	1648	

Fig. 4: Partitioning our data across weeks

As mentioned earlier, different time windows chosen in this project that 1-hour window resulted in more accurate predictions. Fig. 5 illustrates traffic duration for one sample subject in our data set for Monday across three weeks. As it is clear in the plot, there is a pattern in the traffic usage. As mentioned before, next step is to apply the MA technique to smooth the graph. Fig. 6 presents the data after applying the moving average technique. Column MA(24) stands for the computed moving average. It simply calculated by finding the average value for the first 24 windows. Noteworthy, for the first value in column MA(24) we need to find the average for the duration values from time code $t = 1$ to $t = 24$ which is showed by a dotted rectangle and placed for the first value of MA. Noteworthy, since we are averaging 24 values, we put the averaged value around the middle of the window. In this case, we put the first MA value in the time code $t = 13$. Then the second value can be calculated by using the duration values from the time code $t = 2$ to $t = 25$ which is exposed by a dashed rectangle. This value can be placed for the MA at location of time code $t = 14$. Same process can be continued until the last value that find the average from $t = 49$ to $t = 72$ and place it for the time code $t = 61$. Since the number of windows is an even number (24 windows of 1-hour in a day), Centered Moving Average (CMA) should be used to find the actual average value. For computing the CMA, simply average for each pair of MA values can be calculated and placed as the CMA value. For example, for the first value of CMA, we can average first two values of MA and place at the time code of $t = 13$. Similarly, it can be continued until averaging last two values of MA. Fig. 7 presents the duration plot after applying the MA technique.

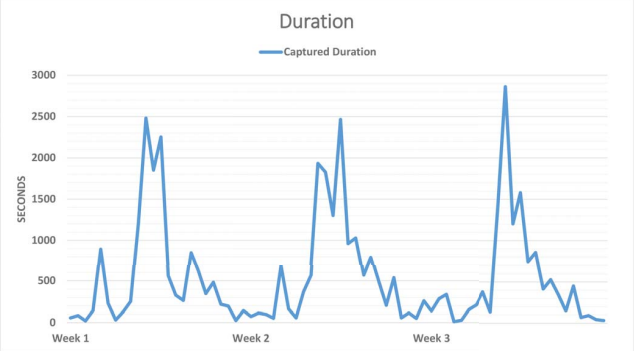

Fig. 5: Internet traffic duration for a sample subject for Monday across three weeks.

Next step is to find the variable containing the seasonality. To do so, seasonality should be extracted from CMA. Average of the division of the original data over the CMA for the first time window across all three weeks can be computed and placed for the first time window across all four weeks as it was stated in the Algorithm 1. The process is summarized in Eq. (3) where m denotes the time window number that starts from 1 to 24. Each time, it calculates for a specific time window and place the result for that specific time window across all four weeks. As the result, seasonal component is extracted from the data.

$$S = [\frac{L[m]}{CMA[m]} + \frac{L[m+24]}{CMA[m+24]} + \frac{L[m+48]}{CMA[m+48]}]/3 \tag{3}$$

Next step is to calculate the deseasonalized value by dividing the original traffic by the seasonal component that computed in the previous step. To finish the process, trend component is also need to be extracted. Linear regression can be applied to find the trend as it is shown in Eq. (4). Where Y is the deseasonalized value that computed earlier and X is the explanatory variable which in this project, time code of t is considered as the explanatory variable. Outputs from the regression model are a and b which they are intercept and slop, respectively. Trend can be computed based on these values in addition to the time code of t, based on Eq. (5). Finally, components of seasonality and trend which are needed for the prediction are founded with the previous steps. Simply by multiplying seasonality and trend, the predicted Internet traffic can be computed.

$$Y = a + bX \tag{4}$$

$$T_t = a + (b \times t) \tag{5}$$

As mentioned earlier, seasonality is the same for all weeks. So, simply it can be placed for the fourth week in our month long data set. Also, trend component can be computed

t	Week	Day	Window	Duration	MA(24)	CMA(24)	S_t	Deseasonalize	T_t	Forecast
1			1	5			0.18298	311.503	594.906	
2			2	83			0.34792	238.561	592.736	
3			3	21			0.38016	55.240	590.566	
4			4	143			0.05227	0.000	588.396	
5			5	890			0.57016	1567.980	586.226	
6			6	229			0.27843	822.481	584.056	
7			7	33			0.24489	0.000	581.886	
8			8	127			0.65839	192.895	579.716	
9			9	254			0.59132	429.547	577.546	
10			10	1207			2.91170	414.535	575.375	
11			11	2480			4.22243	587.339	573.205	
12			12	1851			2.55997	723.055	571.035	
13	Week 1	Monday	13	2251	569.375	569.6875	3.66278	614.834	568.865	
14			14	561	570	570.6458333	1.32657	422.895	566.695	
15			15	328	571.2916667	572.875	1.30628	251.094	564.525	
16			16	265	574.4583333	572.5625	0.73926	358.465	562.355	
17			17	849	570.6666667	566.2291667	1.29441	655.895	560.185	
18			18	622	561.7916667	560.4166667	0.87568	710.307	558.015	
19			19	346	559.0416667	559.5416667	0.41756	0.000	555.845	
20			20	481	560.0416667	565.0625	0.87527	0.000	553.675	
21			21	218	570.0833333	576.625	0.19613	0.000	551.504	
22			22	197	583.1666667	598.3125	0.23120	852.061	549.334	
23			23	25	613.4583333	599.8541667	0.06755	370.106	547.164	
24			24	143	586.25	574.8125	0.24787	576.908	544.994	
25			1	72	563.375	567.8125	0.18298	393.477	542.824	
26			2	114	572.25	580.5833333	0.34792	327.662	540.654	
27			3	97	588.9166667	603.5208333	0.38016	255.158	538.484	
28			4	52	618.125	624.4375	0.05227	0.000	536.314	
29			5	681	630.75	629.5833333	0.57016	1194.401	534.144	
30			6	163	628.4166667	625.75	0.27843	585.434	531.974	
31			7	56	623.0833333	620.1666667	0.24489	0.000	529.803	
32			8	368	617.25	618.4166667	0.65839	558.941	527.633	
33			9	568	619.5833333	616.2083333	0.59132	960.561	525.463	
34			10	1934	612.8333333	611.125	2.91170	664.217	523.293	
35			11	1827	609.4166667	609.9375	4.22243	432.689	521.123	
36	Week 2	Monday	12	1302	610.4583333	612.875	2.55997	508.600	518.953	
37			13	2465	615.2916667	616.6666667	3.66278	672.987	516.783	
38			14	961	618.0416667	621.6041667	1.32657	724.425	514.613	
39			15	1029	625.1666667	630.1458333	1.30628	787.731	512.443	
40			16	568	635.125	634.2916667	0.73926	768.332	510.273	
41			17	793	633.4583333	619.8541667	1.29441	612.632	508.102	
42			18	494	606.25	606.125	0.87568	564.134	505.932	
43			19	206	606	609.25	0.41756	0.000	503.762	
44			20	537	612.5	612.7708333	0.87527	0.000	501.592	
45			21	56	613.0416667	604	0.19613	0.000	499.422	
46			22	115	594.9583333	583.875	0.23120	497.396	497.252	
47			23	50	572.7916667	594.3333333	0.06755	740.213	495.082	
48			24	259	615.875	613.8125	0.24787	1044.890	492.912	
49			1	138	611.75	593.3333333	0.18298	754.164	490.742	
50			2	285	574.9166667	570.3541667	0.34792	819.154	488.572	
51			3	336	565.7916667	562.1875	0.38016	883.846	486.402	
52			4	12	558.5833333	555.3958333	0.05227	0.000	484.231	
53			5	28	552.2083333	546.7083333	0.57016	49.109	482.061	
54			6	157	541.2083333	538.1875	0.27843	563.885	479.891	
55			7	212	535.1666667	533.9791667	0.24489	0.000	477.721	
56			8	381	532.7916667	531.0208333	0.65839	578.686	475.551	
57			9	134	529.25	529.4791667	0.59132	226.611	473.381	
58			10	1402	529.7083333	529.2083333	2.91170	481.506	471.211	
59			11	2861	528.7083333	528.5416667	4.22243	677.571	469.041	
60	Week 3	Monday	12	1203	528.375	523.625	2.55997	469.927	466.871	
61			13	1581	518.875		3.66278	431.640	464.701	
62			14	742			1.32657	559.337	462.530	
63			15	856			1.30628	655.294	460.360	
64			16	415			0.73926	561.369	458.190	
65			17	529			1.29441	408.679	456.020	
66			18	349			0.87568	398.548	453.850	
67			19	149			0.41756	0.000	451.680	
68			20	452			0.87527	0.000	449.510	
69			21	67			0.19613	0.000	447.340	
70			22	91			0.23120	393.592	445.170	
71			23	42			0.06755	621.779	443.000	
72			24	31			0.24787	125.064	440.830	
73			1	36			0.18298		438.659	80
74			2	51			0.34792		436.489	152
75			3	143			0.38016		434.319	165
76			4	5			0.05227		432.149	23
77			5	357			0.57016		429.979	245
78			6	412			0.27843		427.809	119
79			7	429			0.24489		425.639	104
80			8	283			0.65839		423.469	279
81			9	468			0.59132		421.299	249
82			10	1495			2.91170		419.129	1220
83			11	2237			4.22243		416.958	1761
84	Week 4	Monday	12	851			2.55997		414.788	1062
85			13	1612			3.66278		412.618	1511
86			14	605			1.32657		410.448	544
87			15	593			1.30628		408.278	533
88			16	249			0.73926		406.108	300
89			17	256			1.29441		403.938	523
90			18	293			0.87568		401.768	352
91			19	307			0.41756		399.598	167
92			20	466			0.87527		397.428	348
93			21	53			0.19613		395.258	78
94			22	47			0.23120		393.087	91
95			23	12			0.06755		390.917	26
96			24	37			0.24787		388.747	96

Fig. 6: Prediction framework applied to first three weeks of traffic and predict the fourth week.

based on the variables a, b and t that are already computed. Ultimately, multiplication of seasonality and trend can be

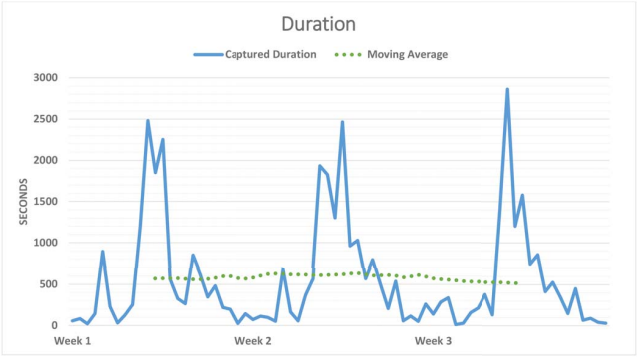

Fig. 7: Internet traffic duration for a sample subject for Monday after applying the moving average technique.

calculated and the fourth week of data can be predicted. This prediction is also presented in the Fig. 6 and labeled as $Forecast$.

Fig. 8 plots the predicted duration data for the fourth week. Furthermore, Fig. 9 plots the predicted data compared to the captured data. Same procedure can be applied to any other day of week or even can be applied to predict short-term prediction like a week or month in advance or even long-term prediction like a year or more.

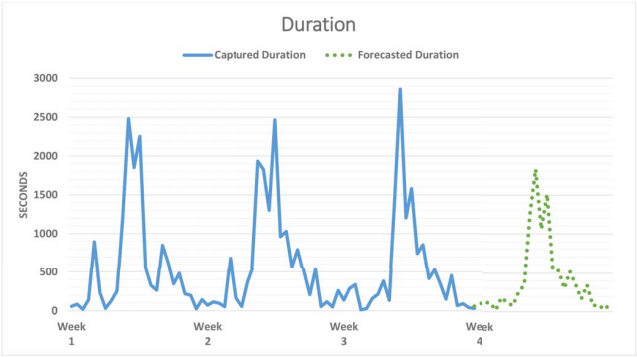

Fig. 8: Predicted Internet traffic duration for a sample subject a for Monday.

Recall from previous discussions, octets can be employed to predict the future data. Similar procedure that was described in Algorithm 1 can be used by considering traffic volume. Fig. 10 illustrates a plot which is comparing the predicted volume to the captured traffic volume for one sample subject for Monday.

V. Results of Data Prediction on Our Data Sets

In this section we present results of applying our forecasting technique to predict future data. The time window for the prediction across the subjects was chosen in a range of 15-second to 24-hour. We observed that a time window of 1-hour can result in more accurate prediction. Due to the space limitations, presenting all the possible predictions for all time windows is not possible. We just present results for the 1-hour time window in this paper, that gave us best accuracy in

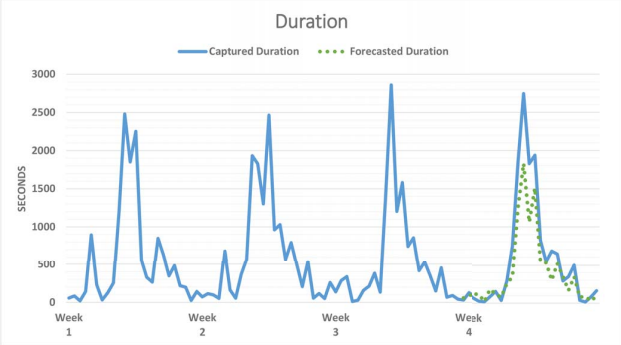

Fig. 9: Predicted versus captured Internet traffic duration for a sample subject for Monday.

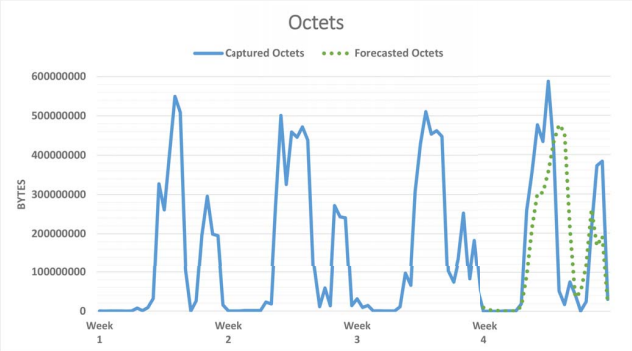

Fig. 10: Predicted versus captured Internet traffic volume for a sample subject for Monday.

forecasting both volume and duration. In Fig. 11, we present the average error in volume prediction across multiple time windows, where we see for all weeks used to forecast, the window of 1-hour gives us the lowest error. The result is the same for forecasting duration also. As such, we report results later in this paper for the 1-hour profiling window only. Also, instead of presenting forecasting results for each day of the week, we present summaries only due to space limitations.

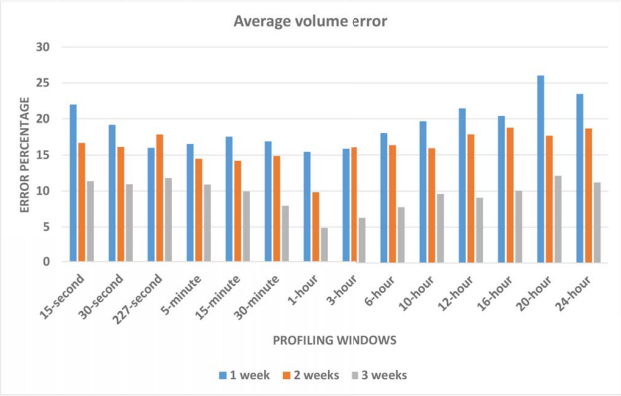

Fig. 11: Average volume error across different time windows.

Fig. 12 summarizes our results. The X-axis is presenting 66 subjects labeled $S_1, S_2, S_3, ..., S_{66}$. Three bars are presented for each user. The left bar is employed to show the average duration error based on 1 week of data. Middle bar presents the average duration error based on 2 weeks of data and finally right bar is the representation for average duration error based on the previous 3 weeks of data. In addition, Y-axis denotes the average duration error in percentage that can be calculated by using the Eq. (6). Data reported are averaged for prediction across all weekdays in a week.

$$percent\ error = \frac{|predicted\ value - captured\ value|}{captured\ value} \times 100\% \quad (6)$$

Same procedure as employed for duration, was used by considering variable octets. Similarly, Fig. 13 summarizes the average error in predicting the future volume of Internet usage for all the subjects in our data set. Again, averages across all weekdays are presented. As it is clear, the framework is predicting effectively since it can predict the future volume of usage based on last three weeks of data with an average error of 4.86%.

The results presented here convince us that our proposed technique can effectively forecast duration and volume of Internet usage from past profiles, with good accuracy depending on profiling window chosen. Also, many other variables can be extracted from a NetFlow data like Destination IP addresses, Ports and Protocols that do provide information that is potentially useful for forecasting. However, in this study we only focused on duration and volume of Internet usage alone that have minimal exposure from a privacy perspective.

VI. Practical Impact of our Work in this Paper

Demonstrating the feasibility of predicting users Internet traffic based on duration and volume 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 more secure Internet access where usage duration and volume of the incoming traffic can be compared with the predicted usage to detect anomalies and make alerts. It also can be useful in run-time management of network resources via superior forecasting. In addition, it can be helpful for smarter advertising via forecasting usage time. Since, we demonstrated that this framework can predict data with a very low amount of average error, such system will be practical. It is also possible to predict required network resources in advance like a week or couple of weeks in advance. However, for such applications, we need more data samples from many more subjects with more diversity beyond campus environment, 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 environment; incorporating machine learning techniques for data processing and more.

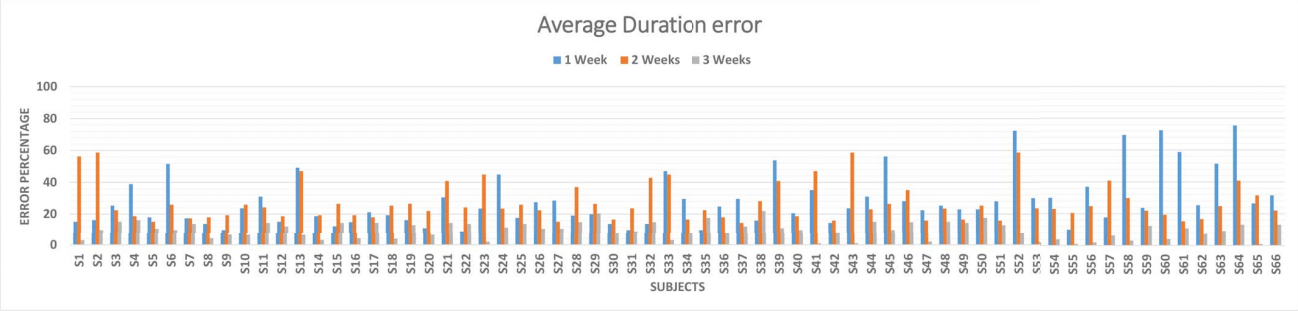

Fig. 12: Error in forecasting duration of Internet usage based on 1, 2 and 3 weeks of data for all the 66 subjects.

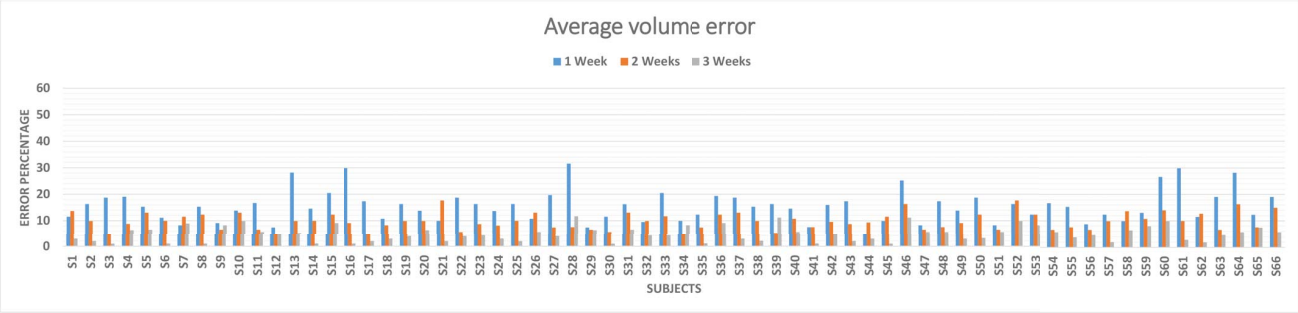

Fig. 13: Error in forecasting volume of Internet usage based on 1, 2 and 3 weeks of data for all the 66 subjects.

VII. Conclusion

In this paper, we demonstrate that duration and volume can act as key variables for predicting Internet traffic. With a sample of college students and privacy preserving NetFlow logs, our prediction framework demonstrate positive results. We do agree that our prediction framework presented here is only a starting point. There are definitely avenues for improvement. For instance, one could consider more features along with duration and volume like destination IP addresses, ports and protocols, which are obtainable from NetFlow data. Also, state-of-the-art machine learning techniques can be used to predict much more accurately. However, there is always a privacy vs. usability trade-off here since with more NetFlow features, accuracy of prediction will improve, but at the cost of privacy. There are all potentially open issues that we believe our work in this paper can inspire.

Acknowledgments

This work was supported in part by the US National Science foundations under grant IIS 1559588 and CNS 1718071. Any opinions, thoughts and findings are those of the authors and do not reflect views of the funding agency.

References

[1] 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.

[2] K. C. Claffy, H.-W. Braun, and G. C. Polyzos, "A parameterizable methodology for internet traffic flow profiling," *IEEE Journal on selected areas in communications*, vol. 13, no. 8, pp. 1481–1494, 1995.

[3] Q. Xu, J. Erman, A. Gerber, Z. Mao, J. Pang, and S. Venkataraman, "Identifying diverse usage behaviors of smartphone apps," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM, 2011, pp. 329–344.

[4] 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.

[5] F. H. Montgomery, S. Chellappan, R. Kotikalapudi, D. C. Wunsch, and K. F. Lutzen, "Monitoring student internet patterns: Big brother or promoting mental health?" *Journal of Technology in Human Services*, vol. 31, no. 1, pp. 61–70, 2013.

[6] L. Malott, S. P. Vishwanathan, and S. Chellappan, "Differences in internet usage patterns with stress and anxiety among college students," in *e-Health Networking, Applications & Services (Healthcom), 2013 IEEE 15th International Conference on*. IEEE, 2013, pp. 664–668.

[7] S. P. Vishwanathan, L. Malott, S. Chellappan, and P. M. Doraiswamy, "An empirical study on symptoms of heavier internet usage among young adults," in *Advanced Networks and Telecommuncations Systems (ANTS), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1–6.

[8] S. Sarmadi, M. Li, and S. Chellappan, "On the feasibility of profiling internet users based on volume and time of usage," in *Communications (LATINCOM), 2017 IEEE 9th Latin-American Conference on*. IEEE, 2017, pp. 1–6.

[9] A. E. Schlosser, S. Shavitt, and A. Kanfer, "Survey of internet users attitudes toward internet advertising," *Journal of interactive marketing*, vol. 13, no. 3, pp. 34–54, 1999.

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

[11] C. Chatfield, *Time-series forecasting*. CRC Press, 2000.

[12] J. Grandell, "Time series analysis," *Lecture Notes (KTH, Sweden, 2000). http:// www. math. kth. se/ matstat/ gru/ sf2943/ ts. pdf*, 1998.