2016 30th International Conference on Advanced Information Networking and Applications Workshops

Traffic Profiling: Evaluating Stability in Multi-Device User Environments

Taimur Bakhshi, Bogdan Ghita
Center for Security Communications and Network Research
University of Plymouth
Plymouth, United Kingdom
{taimur.bakhshi, bogdan.ghita}@plymouth.ac.uk

Abstract— **An ever-growing number of network connected devices per user and changing usage preferences in current networks require novel monitoring techniques be put in place by operators to identify and, if possible, predict user requirements and behaviour. This study utilises NetFlow and cluster analysis to derive six unique traffic profiles namely, 'communicators', 'concealers', 'downloaders', 'all-rounders' and 'high' and 'low intensity web surfers', of around two hundred users (each having multiple devices) based on their application usage trends. During 4 weeks of observation, the analysis indicates that profiles are rather static, with the average probability for a specific device to change its profile ranging between 3-19% over any consecutive 24 hours. Additionally, user devices exhibited a greater probability (up to 8%) of switching to profiles that were similar in terms of application usage ratios due to proportional variations in the same user activity rather than any major usage changes. User devices in some profiles such as 'concealers' and 'downloaders', showed minimal change in application proportion attributed to dedicated P2P and FTP usage. The reported high level of consistency in all profile memberships led to recommending the use of resulting profile baselines for effective monitoring, policy enforcement and anomaly detection in modern networks.**

Keywords- network management; network monitoring; user traffic profiling

I. INTRODUCTION

User traffic characterization is of fundamental importance in modern networking environments for understanding user behaviour which assists administrators and service providers in traffic optimization, capacity planning and improving security. Analysing network workload usually requires collecting statistics around multiple traffic features ranging from total traffic volume, number and duration of user connections to application usage per consumer and provides network managers the ability to anticipate both business and technological updates [1-6]. Proliferation of high speed residential broadband access over recent years, coupled with the growing number of devices per household have compelled providers to seek ways of modelling and understanding user behaviour for improved service differentiation and context based charging [2][3]. Segregation of users into traffic classes or profiles as per their application usage makes this task easier. User traffic profiling in present residential and enterprise networks however, is no longer limited to a single device but multiple devices and variation in user activities per device makes profiling overall subscriber behaviour even more interesting. Effective network management, particularly in modern networks, where subscribers have more than one device sharing a common internet connection, requires an understanding of traffic patterns both inside the network as well as externally [2][6]. Modelling user behaviour in present networks, therefore, remains a critical field of investigation and evaluating the frequency of profile transitions among multiple devices per user may give additional insight into user behaviour having applicability in a broad range of avenues from network management to security. This paper focuses on evaluating traffic profiling as a means of continuous network monitoring by studying user activities in a residential network consisting of two hundred subscriber premises, each with multiple devices, over a four-week timeframe. The methodology follows 1) grouping popular internet applications into distinct categories and classifying traffic using destination web server IP addresses and port numbers 2) developing device traffic profiles based on application usage trends using unsupervised cluster analysis 3) evaluating the probability of inter-profile changes among devices per premises to establish transition trends and re-profiling frequency and 4) utilizing traffic profiles for continued effective monitoring of the network. The rest of this paper is organized as follows. Section II presents a background on related work in user behaviour characterization, highlights challenges in traffic classification and details clustering techniques. Section III presents the methodology used for data collection and feature based clustering. Section IV evaluates the resulting traffic profiles and makes recommendations towards continuous effective monitoring of users. Section V draws final conclusions.

II. BACKGROUND

The predominant method of characterising user behaviour in both enterprise and residential networks has been to cluster users based on peculiarities in traffic features, using flow and packet based measurements or report these as standalone attributes for monitoring overall network traffic as highlighted in [2], [4], [5] and [6]. More recently, the proliferation of high speed residential broadband and increase in the number of devices per household has highlighted the need to understand device traffic from inside the network for effective network management and improved

978-1-5090-2461-2/16 $31.00 © 2016 IEEE
DOI 10.1109/WAINA.2016.8

security. For example, Xu. et. al used port number affiliations as the primary feature for clustering device traffic inside user residences and identifying internet malware [2]. Traffic profiling is hence, no longer isolated to a single device per user and in modern networks individual device profiling for users with multiple devices presents an interesting avenue for understanding the temporal nature of user activity per device. This may further aid in ascertaining if device traffic profiling lends considerable consistency over time to serve as an effective tool for network monitoring and management, either internally by subscribers themselves, their network administrators (in enterprise networks) or externally by service providers. With regards to application usage/identification for traffic profiling, port assignments remain the prevalent method in prior studies owing the inherent challenges in traffic classification. Novel methods such as machine learning classification algorithms or deep packet inspection techniques either involve substantial processing overhead or highly sanitized and pre-processed records for getting meaningful results as detailed in [7], [8], [9] and [10]. Application traffic classification is hence a research problem on its own. Service providers and network administrator have an imminent need for extrapolating subscriber application usage and rely on commodity tools like [11] and [12] which partially circumvent the traffic identification problem by including pre-defined customisable webserver IP addresses of frequently used applications and websites, matching these to user requests for accounting. This scheme may seem limited but with careful planning and continuous updating can be effectively used to report top applications and website visitations. To address scalability issues, the present study implemented a similar approach for identifying application traffic detailed later. The next section discusses the methodology used for user traffic profiling.

III. METHODOLOGY

The present study aimed at developing a novel traffic monitoring framework in residential networks through user traffic profiling based on users' application trends. The frequency of change in user activity was further assessed by observing profile transitions per user-device to evaluate if derived profiles granted significant stability over time to serve as benchmarks for formulating network management policies. The following subsections detail the methodology employed for profile derivation through cluster analysis, application identification and data collection.

A. Traffic profiling

The present work required partitioning user devices into groups based on proportional variations in application usage to derive traffic profiles. For this purpose two prominent un-supervised clustering techniques were employed a) hierarchical agglomerative clustering and b) k-means [13][14]. Hierarchical clustering puts each observation in its own cluster and then calculates the distances between all observations, pairing the closest two in a recurring fashion. K-means on the other hand, minimizes a given number of vectors by choosing k random vectors as initial cluster centers and assigning each vector to a cluster as determined by distance metric comparison with the cluster center (a squared error function) as given in (1). Cluster centers are then recomputed as the average of the cluster members. This iteration continues repeatedly, ending either when the clusters converge or a specified number of iterations have passed [17].

$$J = \Sigma_{j=(1,k)} \Sigma_{i=(1,n)} ||x_{ij} - c_j||^2 \quad (1)$$

In (1) above, c_j represents the cluster center, n equals the size of the sample space and k is the chosen value for number of unique clusters (centroids). Hence, using k-means, n entities, translating for user devices in the present case, can be partitioned into k groups or profiles. Value of k is of significant importance as it directly influences the number of traffic profiles and affects over-fitting of users into profiles. The next subsection discusses the scheme used for identification and grouping of applications per device for subsequent profiling.

B. Application identification

User traffic was classified by matching user requests (NetFlow records) against destination IP addresses and ports used by popular internet applications. These were further cross-referenced against commercial tools such as NetFlow Analyser [11] and PRTG Network Monitor [12] for greater accuracy. To account for replication in nature of user activities, applications were further grouped into distinct categories as depicted in Table 1. A unique website visitation or application usage on user device could therefore be defined by the vector u_{ij} given in (2).

$$u_{ij} = [w_{ij}, e_{ij}, d_{ij}, v_{ij}, g_{ij}, c_{ij}, t_{ij}, z_{ij}] \quad (2)$$

In equation (2) above, i and j are unique per user and user device respectively and remaining entities represent the application usage percentage in accordance with Table 1.

TABLE I. APPLICATOIN GROUPS

Application Tier	Website, Destination Port
Web browsing (w)	Web browsing: http(s) except below groups
Emailing (e)	Gmail, Ymail, AOL, Hotmail, Outlook, SMTP, POP3, IMAP
Downloading (d)	BitTorrent, VUZE Torrent, FTP
Video streaming (v)	YouTube, Netflix, Lovefilm, Megavideo, Metacafe
Games (g)	n4g, uk-ign, freelotto, 8-ball pool
Communication (c)	Skype, Net2Phone, MSN Messenger, Yahoo Messenger, GTalk
Uknown traffic (t)	Unaccounted TCP and UDP traffic
Network utility (z)	DNS queries, Network Multicasts

C. Monitoring setup

The study used NetFlow records exported from the default gateway of a residential complex housing approximately two hundred users over a span of four weeks from 01/02/2015 to 28/02/2015, with all user premises connected via dedicated ports on access switches to the central router as shown in Fig 1. Users had own CPEs such as home routers for connecting multiple devices to the internet. SNMP instances on access switches reported IP addresses of devices connected per premises to the central monitoring platform.

D. Data collection and pre-processing

NetFlow records collected by the central collector were concatenated and customised every 24 hours to build bidirectional flow records incorporating traffic statistics. The resulting logs were processed to quantize user device activity (flows) as a percentage of application usage in accordance with Table 1. Afterwards, SNMP monitoring information from access switches was used to associate individual devices to user premises. Network activity for a user device $u_{1,1}$ on a specific 24hour interval, e.g. [01/02/2015], could therefore be represented by the application distribution vector given in (3).

$$u_{1,1}\,[01/02/2015] = [83\ 0.5\ 1.7\ 1.8\ 2\ 0.1\ 0.7\ 9.9] \qquad (3)$$

Once application distribution vectors per user device were collected, traffic profiling was done using both agglomerative hierarchical clustering as well as Hartigan and Wong implementation of k-means [18], using R. The resulting traffic profiles and associated analysis are described in the following section.

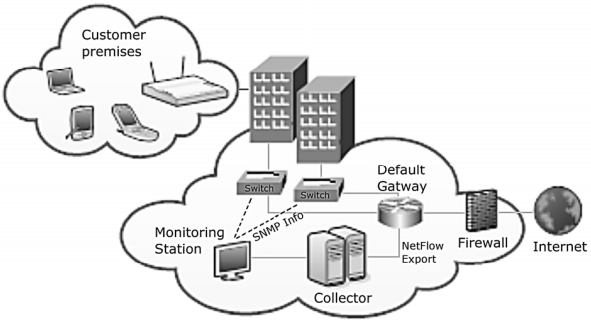

Figure 1. Network monitoring setup

IV. Evaluation

A. Cluster analysis

A total of 10095 unique traffic distribution vectors for user devices were examined comprising approximately 178.21 million flows. The resulting vectors were subjected to hierarchical and k-means clustering and cluster memberships were evaluated for cluster sizes ranging between 2 to 10 as depicted in Table 2. Using the default maximum distance linkage (or complete) method in hierarchical clustering, profile membership numbers resulted in minimal number of observations in some clusters. The more advanced Ward's method and k-means, however, resulted in much reasonable membership numbers across all the derived clusters. The next step was finding the optimal number of clusters (translating for traffic profiles) which would appropriately reflect user activities. Clusters derived using k-means were examined starting from k=2, using Everitt and Hothorn technique given in [19]. This technique aims at finding the curve in plot of 'within groups sum of squares distance' per observation in each cluster against k for suggesting an appropriate number of profiles that fit the input data. The corresponding plot for present data is given in Fig. 2 where a significant drop can be seen up to a cluster size k=6, with minimal variations up to k=15, which indicates an optimal value of 6 profiles for the entire subscriber base. Application distribution ratios for profiles derived using both hierarchical clustering (Ward's method) and k-means were afterwards, compared in parallel to ascertain which set presented meaningful results. While profile membership numbers per cluster using either method were quite similar, profiles derived using k-means gave a much clearer segregation of user activities. For example, for six profiles, hierarchical clustering resulted in three profiles having quite similar web-browsing ratios of 92.26 %, 83.45% and 77.39% compared to only two profiles with high web-browsing and well parted usage ratios of 90.04 % and 62.84 % derived by k-means. This trend continued up until the maximum examined value of ten profiles. Hence, for the present study the profiles derived using k-means (k=6) represented a meaningful balance catering for both heavy membership profiles as well as lower ones without compromising too much on mutual exclusivity or overfitting.

TABLE II. Profile Membership Distribution Per Cluster

k	hclust (method= max)	hclust (method=ward)	k-means
2	10094, 1	6316, 3779	8236, 1859
3	10002, 92, 1	6316, 2365, 1414	6172, 2576, 1347
4	9742, 260, 92, 1	6316, 2365, 1216, 198	5013, 2642, 1272, 1168
5	9742, 255, 92, 5, 1	3904, 2365, 2412, 1216 , 198	6339, 2473, 970, 228, 85
6	8815, 927, 255, 92, 5, 1	3904, 2412, 2365, 687, 529, 198	5013, 2644, 1249, 880, 224, 85
7	7868, 947, 920, 255, 92, 5, 1	3904, 2412, 1253, 1112, 687, 529, 198	3965, 2660, 1538, 828, 224, 795, 85
8	7868, 947, 920, 255, 90, 7, 5, 1	3904, 2412, 1253, 1022, 687, 529, 198, 90	3736, 2607, 1574, 765, 693, 417, 218, 85
9	7868, 947, 920, 255, 90, 7, 5, 2, 1	3904, 2412, 1253, 757, 687, 529 , 265, 198, 90	4318, 2637, 944, 699, 602, 439, 222, 180, 54
10	7868, 920, 905, 255, 90, 42, 7, 5, 2, 1	2412, 2186, 1718, 1253, 757, 687, 529, 265, 198, 90	3623, 2399, 1018, 856, 667, 626, 359, 248, 214, 85

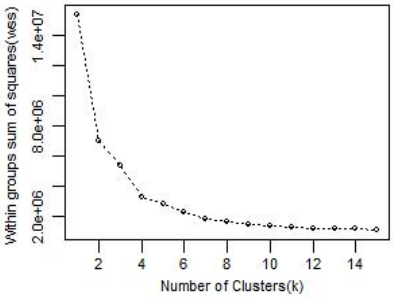

Figure 2. Identifying correct number of clusters (*wss vs. k*)

B. Results

Since user traffic profiling was done on the basis of application usage, IP addresses and flows were excluded while clustering. Also general network service traffic (z_{ij}) such as DNS queries are not a user-triggered application but a functional one, hence, it was also excluded while clustering users and later separately calculated as a percentage of total network flows generated per profile. The resulting traffic profiles (k=6), detailing the application traffic distribution among user traffic profiles are given in Table. 3, with detailed traffic statistics per profile given in Figs 3 – 6. From a practical perspective the resulting profiles showed a great deal of variation in user activity i.e. their application usage. For example, some users concentrated on web-browsing only with minimal use of other applications such as video streaming or playing online games. We categorized user profiles accordingly with respect to user activities. For example, Profile 1 concentrated on web browsing with minimal usage of other applications and had the highest number of devices and users and was therefore, labelled as high intensity web-surfers. Profile 2 also concentrated on web browsing; however, browsing was significantly less (68.92%) as compared to high intensity surfers, hence named low-intensity web surfers. Profile 3 also inclined towards web browsing, but traffic distribution among other activities such as emails, downloads, streaming and games was slightly higher than both high and low intensity surfers. Due to their proportional use of many applications we categorized these profile devices under all-rounders. User devices in Profile 4 heavily tilted towards using communication related applications (88.01%), with negligible traffic in any other tier and were therefore, labelled as communicators. Unknown or concealed traffic accounted for most of Profile 5 at 66.92%. The traffic identification scheme discussed earlier fell considerably short of identifying the applications or website visitations for devices in this profile. Closer analysis of source and destination ports revealed that concealing devices were randomly using un-assigned ports with the majority of traffic attributed to P2P applications. Devices in this profile were categorized as concealers. Due to the low percentage of unidentified application traffic in other profiles in comparison to concealers, unknown network traffic did not significantly influence the overall results (profiles). Lastly, profile 6 mainly focused on downloading data (83.02%) from the internet using FTP and other download applications (torrents) and had the lowest number of devices and users and accordingly labelled as downloaders. Hence, each device per user premises represented a significant discrimination towards a certain mix of user activity. Next we evaluate the consistency in these traffic profiles to comprehend changes in user behaviour in terms of their respective device usage.

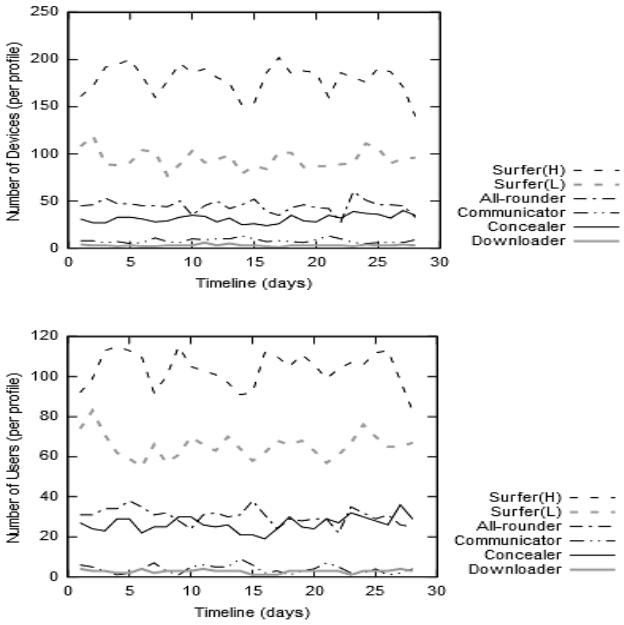

Figure 3. (a) Number of devices and (b) Users per traffic profile

TABLE III. TRAFFIC PROFILES

App. Tiers	Profile 1 (High Intensity Surfers)	Profile 2 (Low Intensity Surfers)	Profile 3 (All-rounders)	Profile 4 (Communicators)	Profile 5 (Concealers)	Profile 6 (Downloaders)
Browsing	90.04	68.92	42.40	3.21	9.63	2.95
Emails	0.75	1.27	2.19	0.25	3.64	2.06
Downloading	1.75	3.46	3.59	0.15	14.44	83.02
Streaming	0.69	1.29	1.60	4.51	0.84	0.05
Games	0.22	0.46	2.19	1.34	0.95	1.45
Communication	0.64	1.67	3.28	88.01	2.17	0.15
Unknown	2.06	8.68	23.25	1.94	66.92	7.80
Net. Util.	3.85	14.25	21.50	0.59	1.41	2.52

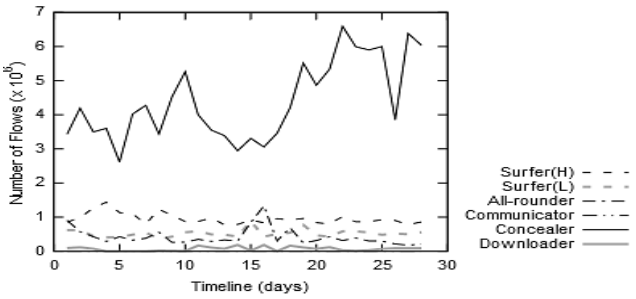

Figure 4. Total flows per traffic profile

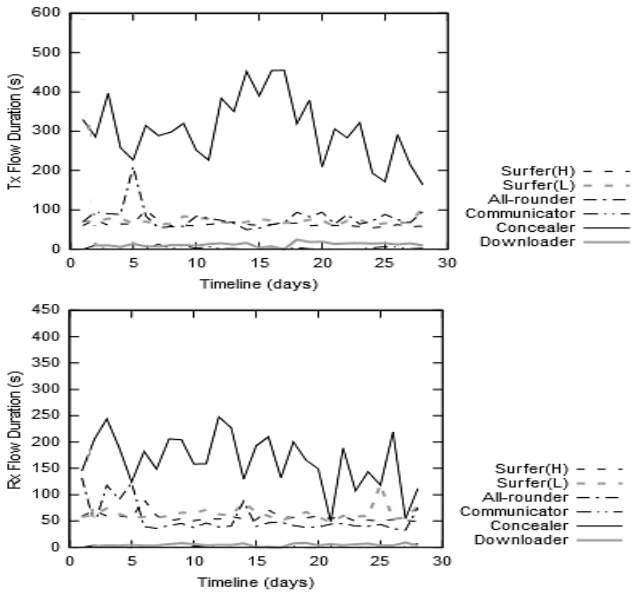

Figure 5. Average duration of (a) transmitted and (b) received flows

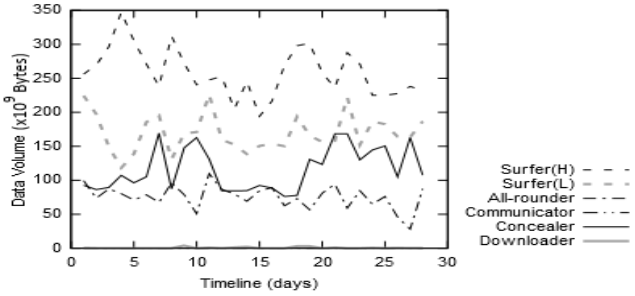

Figure 6. Total data volume per traffic profile

C. *Profile consistency*

Profile consistency emphasises the importance of gaining a better insight to changes in user activity per device as well as requirements or frequency for re-profiling. Fig. 7 shows the Pearson's correlation coefficient giving the number of devices per profile and their corresponding correlations with others [17]. Negative correlation co-efficient refers to an inverse relationship with one profile gaining more devices and the other losing, however, not necessarily among the same profile pairs. Positive values refer to an increase in devices for both pairs. Values closer to zero represent no meaningful relationship, translating for minimal increase or decrease in devices per profile pair. As we can see in Fig. 7, there is a blend of both negative and positive correlations representing changes in number of devices per profile. The average value of correlation co-efficient was further calculated to be -0.0931 by using Fisher's transformation [18]. Being, close to zero, this average represented no significant change in device numbers per profile with a slight bias towards an inverse relationship between each profile pair. To further evaluate user device profile retention, the average probability of change in device profiles per subsequent day of study was computed and is given in Table 4. Downloaders showed the highest consistency in retaining profiles at 97% while all-rounders showed lowest at 81%. The probability of a profile gaining or losing a device every 24 hours is also highlighted in Table 4. Downloaders had the highest probability of gaining a device (60%) with low intensity web surfers having highest probability of losing a device (59%). Where devices did change profiles, the average probabilities of inter-profile transitions every 24 hours are given in Table 5. It was observed that where devices transitioned to a different profile, it was always to profiles having somewhat similar application usage ratios to their own. For example, high surfers would transition to low surfers (8%) compared to other profiles. Concealers did not show any significant change in profile, except to all-rounders which was also minimal (5%). This further emphasized the fact that majority of devices within this group more closely followed an application dictated pattern of behaviour mainly due to P2P activities. Downloaders seldom changed profiles highlighting that the small number of devices in this profile were also being dedicatedly used for downloading files from the internet. Hence, where transition in traffic profiles was observed, it was only due to variation in the same user activity rather than a complete change of role per device. Users therefore, continued to use the same devices for same kind of activity albeit in varying proportions rather than showing drastic changes in their normal routine.

D. *Network monitoring*

Due to the high levels of consistency reported in membership profiles, once traffic profiles have been derived based on application usage ratios, baselines of network traffic per profile depicted in Figs. 4-6 provide an intuitive means to monitor the network. Daily aggregate traffic can be effectively examined by analysing value changes in traffic attributes per profile with any abnormalities serving as an advisory to trigger a re-evaluation of profiles and also identify network anomalies. For end users wanting to better manage their data usage, service providers may employ traffic profiling to place users into correct subscription models while also providing them with their daily traffic projections through service provider portals or customer home gateways. Traffic profiles provide administrators with enhanced capability to monitor network activity and update capacity based on anticipated user behaviour for achieving

better quality of experience. It may also aid in protecting users from security threats or in enforcing policies. For example, in the present case concealers (P2P users) could either be rate-limited or blocked by making provisions in the firewall shown in Fig.1 to enforce the underlying network usage policy.

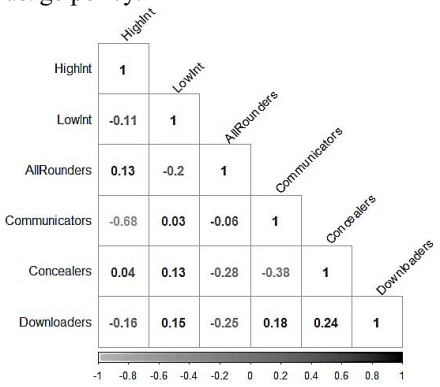

Figure 7. Pearson Correlation Co-efficient between Device Profiles

V. CONCLUSION & FUTURE WORK

The present work focused on profiling multi-device user traffic in a modern residential network housing over two hundred user premises based on application usage using cluster analysis. Six unique user profiles were derived and benchmarked every twenty four hours to ascertain their stability. Over the 4week observation period, the analysis indicates that the number of users and devices per profile remained fairly consistent. Any inter-profile transitions were mainly due to proportional variation in same kind of user activity triggering a device profile change (to a somewhat similar profile in terms of application usage). Based on the overall high rate of profile consistency reported even in this multi-device per tenant environment, the study made recommendations towards effective network management by using user profiles to define and implement network policies. Our future work will focus on applying the proposed traffic profiling methodology to different networks to further validate the employability of this approach in network monitoring and management.

TABLE IV. AVERAGE PROBABILITY OF PROFILE CHANGE (/24 HOURS)

User Profiles	Probability of No Change	Probability of Change		
		Change	***Prob. Device Gain***	***Prob. Device Loss***
H. Int Surfer	0.88	0.11	0.46	0.53
L. Int Surfer	0.87	0.13	0.40	0.59
All-rounder	0.81	0.18	0.42	0.57
Comms.	0.84	0.15	0.57	0.42
Concealers	0.87	0.12	0.48	0.52
Downloaders	0.97	0.03	0.60	0.40

TABLE V. AVERAGE PROBABILITY OF PROFILE CHANGE (/24 HOURS)

User Profiles	High Int.	Low Int.	All-Rnd.	Com.	Conc.	Down.
High Int	0.88	0.08	0.01	0.01	0.008	0.002
Low Int	0.08	0.87	0.03	0.0015	0.007	0.001
All-Rnd.	0.03	0.11	0.82	0.003	0.032	0.0005
Comm.	0.06	0.09	0.05	0.84	0.001	0.001
Conceal.	0.03	0.03	0.05	0.0006	0.87	0.002
Down.	0.001	0.0005	0.0001	0.0004	0.0005	0.97

REFERENCES

[1] Callado A., Kamienski C., Szabo G., Gero B., Kelner J., Fernandes S., Sadok D.,"A Survey on Internet Traffic Identification," Communications Surveys & Tutorials, IEEE 2009, vol.11, no.3, pp.37-52.

[2] Kuai X., Feng W., Lin G., Jianhua G., Yaohui J., "Characterizing home network traffic: an inside view". Personal and Ubiquitous Computing, 2014, 18(4): pp. 967-975.

[3] Dainotti, A., Pescape A, Claffy K, "Issues and future directions in traffic classification," Network, IEEE 2012, vol.26, no.1, pp.35-40.

[4] Hongbo J., Zihui G., Shudong J., Jia W., "Network prefix-level traffic profiling: Characterizing, modeling, and evaluation". Comput. Netw. 2010, 54, 18, pp.3327-3340.

[5] Humberto T. M. N, Leonardo C. D. R, Pedro H. C. G, Jussara M. A, Wagner M. Jr., Virgilio A. F. A, "Characterizing Broadband User Behavior", NRBC 2004, NY, USA

[6] Jinbang C., Wei Z., Urvoy-Keller, G., "Traffic profiling for modern enterprise networks: A case study," IEEE 20th International Workshop on LAMAN, 2014, vol.1, no.6, pp.1,6, 21-23

[7] Iliofotou M., Gallagher B., Eliassi-Rad T., Xie G., Faloutsos M., "Profiling-By-Association: a resilient traffic profiling solution for the internet backbone". *Proceedings of* Co-NEXT 2010. ACM, NY, USA

[8] Williams N., Zander S., Armitage G., "A preliminary performance comparison of five ML algorithms for practical IP traffic flow classification". SIGCOMM Commun. Rev. 36, 5, 2006, pp.5-16.

[9] Camacho J., Padilla P., García-Teodoro P., Díaz-Verdejo J., "A generalizable dynamic flow pairing method for traffic classification", Computer Networks, Vol 57, Iss 14, 4, 2013, pp.2718-2732.

[10] Bujlow T., Carela-Español V., Barlet-Ros P., "Independent comparison of popular DPI tools for traffic classification", Computer Networks, Volume 76, 15 January 2015, pp.75-89.

[11] Cisco NetFlow Analyzer, Website: http://www.cisco.com/netflow/

[12] PRTG NetMon tool, Website: http://www.paessler.com/prtg/

[13] Murtagh, F. and P. Legendre, "Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?" Journal of Classification, 2014. 31(3): p. 274-295.

[14] MacQueen, J., "Some methods for classification and analysis of multivariate observations". Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, 1967, pp 281-297, Berkeley, University of California Press.

[15] Hartigan, J. A. and Wong, M. A., "A K-means clustering algorithm". *Applied Statistics* 28, 1979, pp.100–108.

[16] Brian S. Everitt and Torsten Hothorn, "A Handbook of Statistical Analyses Using", Boca Raton, FL: Chapman & Hall/CRC, 2006.

[17] Keshav S., "Mathematical foundations of computer networking", Addison Wesley; 1 edition, NY, 15 April 2012.

[18] Faller, A. J., "An Average Correlation Coefficient". Journal of Applied Metereology Vol 20, p 203-205, 1981.