User Profiling from Network Traffic via Novel Application-Level Interactions

Gaseb Alotibi[1]; Nathan Clarke[1,2]; Fudong Li[1] and Steven Furnell[1,2,3]
[1]Centre for Security, Communications and Network Research (CSCAN)
Plymouth University, Plymouth, United Kingdom
[2]Security Research Institute, Edith Cowan University, Western Australia
[3] Center for Research in Information and Cyber Security,
Nelson Mandela Metropolitan University, Port Elizabeth, South Africa
Gaseb.alotibi@plymouth.ac.uk

Abstract - **Insider misuse has become a significant issue for organisations. Traditional information security has focussed upon threats from the outside rather than employees. A wide range of research has been undertaken to develop approaches to detect the insider – often referred to as Data Loss Prevention (DLP) tools. Unfortunately, the fundamental limitation of these tools is that they provide information resolved to IP addresses rather than people. This assumes the IP is static and linkable to an individual, which is often not the case. IPs are increasingly unreliable due to the mobile natural of devices and the dynamic allocation of IP addresses. This paper builds upon prior work to propose and investigate a biometric-based behavioural profile created from a novel feature extraction process that identifies user's application-level interactions (e.g. not simply that they are accessing Facebook but whether they are posting, reading or watching a video) from raw network traffic metadata. It also proceeds to describe various types of user's interactions that can be derived from applications. Validation of the model was conducted by collecting 62 GBs of metadata over a 2 months period from 27 participants. The average results of identifying users at first rank in the top three applications Skype, Hotmail and BBC are scored 98.1%, 96.2% and 81.8% respectively.**

Keywords; user profile, user behavioural, user identification, network forensic, forensic investigator

I. INTRODUCTION

More than 3 billion Internet users have utilized Internet services across the world; and this number is growing continuously [1]. Their daily usage includes (but not limited to) web browsing, entertainment (e.g. watching online videos), communication (e.g. making VoIP calls), finance (e.g. online shopping) and office applications (e.g. Google docs). Indeed, many of the traditional desktop applications such as Office are now found as Internet-based services [2], [3]. According to the Office for National Statistics in the UK (2015), 79% of the UK population has daily access to the Internet [4]. Moreover, the study has found that smartphone and portable computer devices have a high Internet penetration in the UK [4]. In addition, a significant trend exists in the use of mobile application has been experienced [5]. According to the Statistical Portal (2016) the number of mobile applications downloaded has increased on average of 140% per year between 2012 and 2015 [6].

Malicious attackers can utilize a variety of approaches, such as hacking, Denial of Service (DoS), social engineering and malicious software to launch different attacks [7]. Various studies have shown that they are often successful. Right from an individual to an organisation, anyone can be a victim of security attacks. These attacks are on rise in the recent years. According to a report of McAfee 2014, the estimated annual cost of the cyber-crime to the global economy is more than $ 400 billion, including companies and individuals [8]. Another recent study by Verizone has found that 70 organisations around the world were the victims of information security attacks in 2015 [9]. These studies reflect the extent of security threats being faced by various organizations and the individuals; and they also highlight the importance and the need for effective security systems needed for safeguarding the information systems. In assessing these security threats, it is very essential to understand the various types of threats and the efficiency of the solutions available for mitigating or overcoming them. These threats can be analysed based on their occurrence or initialization, which can be categorised in to internal and external security threats. Whilst external threats remain an issue, security controls exist to protect outsiders from getting into systems. They do not however often help with the problem of insider threats. Insider threats have grown to be an increasingly significant problem; this is due to the privilege in which the insider user has comparing with the external attacker. According to Information Security Breaches survey in 2015, insider misuse (e.g. unauthorised access) represents 65% of the security breaches in the large organisations [10]. In addition, newer technologies such as smartphone and tablets have become one of the main driving factors behind the security breaches. In large organisations, smartphone and tablets have caused 15% of security or data breaches, doubling the percentage comparing with the previous year [10].

The existing network-based solutions to insider misuse have relied upon two approaches for detecting and/or preventing insider attacks: packet inspection and flow based approach. They both are used to analyse the network traffic to detect and prevent an attack and have been implemented in different security tools such as Intrusion Detection Systems (IDS), network forensic tools, Security Incident and Event

Management (SIEM) systems [11-13]. Whilst these two approaches have their advantages they both fail to truly profile the individual but rather merely the IP address. With the widespread use of mobile technologies (those IP address is constantly changing every time the device is switched on) through to DHCP, where the IP address of a computer can change every time when it is restarted, it is an increasingly unreliable means of undertaking traffic analysis. Furthermore, with an ever increasing volume of encrypted data, deep packet inspection is becoming less useful [14].

Considering these limitations and with a focus on investigating new approaches to handle such limitations, this paper presents the concept and process of deriving user interactions from raw network traffic through the use of metadata independently to user's IP address. Then creating a user-behavioural-based profile that focussed specifically on the user rather than the machine and evaluate the approach across a range of Internet services is the main theme of this study.

The remainder of the paper is structured in the following manner: section 2 reviews existing methods of network traffic analysis for insider threats; section 3 presents the creation of application-level user interactions from network level data, followed by an experiment to validate to what extent these user activities are discriminative. A discussion of the findings and how it could be used in practise is presented in section 5. Finally, the conclusions and future work are given in section 6.

II. Related Work

The research into network traffic analysis started in the 1990s, with the aim of identifying network related attacks [15-16]. Since then, researchers have begun using this technique to investigate different aspects of the problem, such as network behaviour analysis, traffic prioritization, network optimisation and insider misuse. There are two fundamental approaches that can be used to analyse network traffic and being utilised in the existing tools either in detecting/ preventing insider threats: packet or flow based.

A. Packet Based Network Analysis Approach

The concept of this approach is to perform a bit by bit comparison with predefined signatures of known threats. If there is a similarity between the investigated packets (especially in the payload part) and the threat signature, an attack can be detected. A number of tools (both open source and proprietary) have been developed, including Cain and Abel, TCPDump, Wireshark, Xplico and Microsoft Network Monitor, assisting network security analysts and forensic examiners with an easy analysis of packet information that enable a better understanding of how the attack was formed [17], [18]. However, many limitations have been identified in this approach. One of these is time consuming in analysing the large volumes of data. Another significant issue is the growing use of encryption (e.g. SSL/TLS) within network communications – preventing any analysis of the payload [14]. Subsequently, in order to improve the level of performance and its effectiveness various researchers have proposed different methods to deal with deep packet inspection limitation in order to speed up the process in identifying malicious attacks; but they have limited solutions [12], [19-22].

B. Flow Based Network Analysis Approach

Flow-based approaches seek grouping IP packets passing through an observation point in the computer network within a certain time interval based upon a connection profile. All packets that belong to a specific flow have a set of shared properties. These properties may exist in the header or in other different parts of the packet or both [23]. The advantage of using flows is the vast reduction in data that needs to be analysed in comparison to the packet-based approach. The flow record normally consists of various fields such as the time and date stamps, the IP addresses of the communication source and destination, their port numbers, the length of the total payload, and the type of protocols. The flow is normally generated from the raw traffic by using third party applications, such as NetFlow, SFlow, JFlow and IPFIX [23-26]. These applications perform different tasks focusing on flow based analysis such as traffic monitoring, identifying unauthorised network activity and tracing the source of DoS attacks. Typically, this is performed by analysing the current traffic flow and identifying any abnormality based upon the historical traffic profile. Based upon this theory, many methods and tools have been proposed and devised within the flow based network analysis domain [27].

With increasing network bandwidth and the use of encryption, the flow-based approach became prominent in the market in the area of investigating network traffic issues. In addition, the analysis in large capacity networks is considerably timely and fast. Subsequently, this approach is more efficient in detecting network scans and intrusions, the spreading of malware, and monitoring general network usage. However, whilst the approach solves issues of data volume and encryption, it does not provide additional information on how the user interacts with service but only a service is utilised by an IP address.

Both packet and flow based approaches fundamentally rely upon the IP address as the unique identifier to tag individual. Although, they may successfully achieve certain level of accuracy based on their user identifying mechanism, development of technologies, as previously highlighted, in highly mobile and DHCP-enabled environments, these approaches may not be effective enough in analysing the network traffic. This limitation forces network forensics investigators to examine and analyse larger volumes of raw traffic to identify and correlate misuse, which is a time consuming and expensive activity. Far more useful would be to ask the system to present all traffic belonging to the suspect (or hacked account) and for the system to present this data generated from all devices the individual must be connected to the network through a device (e.g. desktop, smartphone, tablet, and laptop).

III. Deriving Application Level User Interactions From Network Data

For successfully profiling an individual it is important to capture and understand the individual from human-level interactions rather than the machine-to-machine interactions (e.g. network management protocols such as ICMP). Therefore, this approach is focused upon extracting and deriving user based interactions from the raw network data. Intuitively, as users are interacting with Internet-based applications, it should be possible to measure that interaction at the flow/packet level through an understanding of the connection parameters (such as connection type, duration, number of packets, and packet size). The following section explains how user actions can be identified and extracted from network metadata.

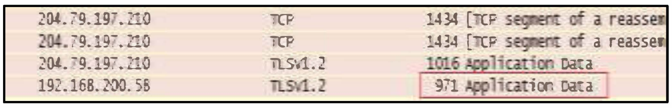
A. Methodology

The approach is based upon the theory that how a user interacts with Internet-based applications on their computer produces a (relatively) unique network packet signature which can subsequently be used to identify the activity.

In order to test this hypothesis, an investigation was undertaken to determine to what extent these signatures could be developed. In the first instance, ten of the most popular internet-based applications were selected for analysis [28]. These applications are Google, YouTube, Skype, Facebook, Dropbox, Hotmail, Twitter, Wikipedia, eBay and BBC. To ensure the resulting analysis was reliable, three researchers were tasked with the collection and analysis of network traces against a predefined set of network captures against user activities (which themselves were repeated 10 times in order to allow for any variance in the resulting network traffic). In our previous paper, an early analysis of these interactions were published which identified that it was possible to determine user interactions with applications through low level network data [29], [30]. The next sections explain the process in determining these signatures in some applications

1) TCP Protocol Signature

TCP protocol is one of the main protocols utilised in the computer network. Usually three forms of signatures exist in this protocol. First one is 'one packet signature'. Fig. 1 shows the patterns when a recipient is added or when a new email button is clicked while using the Hotmail. When the recipient is added or new email button is clicked by a user, one packet is sent from Hotmail server to the client with size 971 bytes.



204.79.197.210	TCP	1434 [TCP segment of a reassem
204.79.197.210	TCP	1434 [TCP segment of a reassem
204.79.197.210	TLSv1.2	1016 Application Data
192.168.200.56	TLSv1.2	971 Application Data

Figure 1. Add recipient on Hotmail

The second type of signature is 'Multiple Packets signature'. When the user starts typing, while chatting on Facebook, 2 packets are sent from the client to a Facebook server. The total size of two of these packets is 1,502 bytes (i.e. 1434+68 or 1169+333). These packets are sent in less than one millisecond timeframe as shown in Fig. 2.

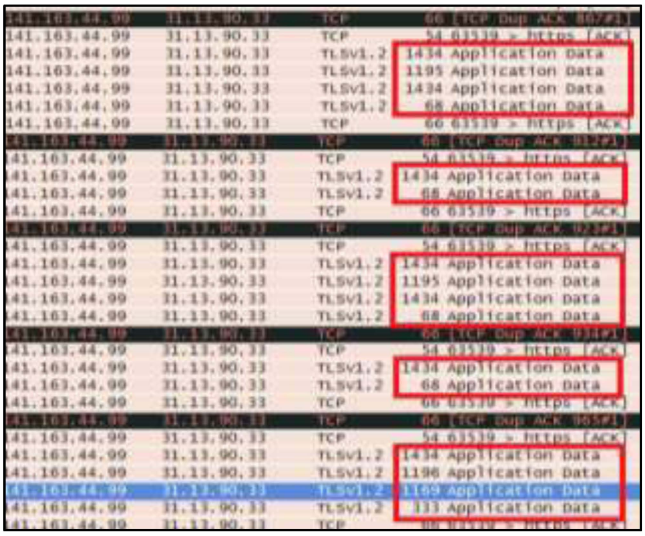
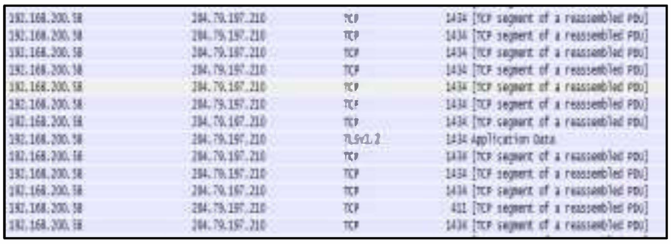


Figure 2. Typing on Facebook

The third type is 'stream packets'. Attaching a file in the Hotmail is an example that leads to the initiation of such packets. When the file is attached, stream of packets starts moving from client to Hotmail web server with maximum transmission units (MTU) as shown in Fig. 3. This process would depend on the file size and would continue till the whole file is attached.

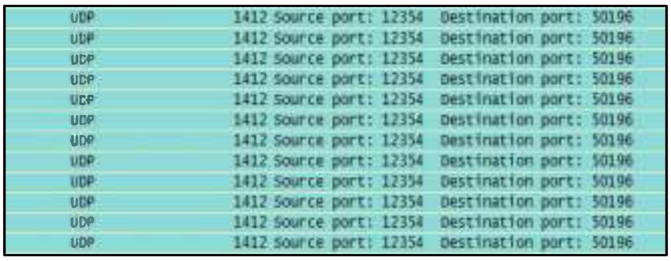


192.168.200.58	204.79.197.210	TCP	1434 [TCP segment of a reassembled PDU]
192.168.200.58	204.79.197.210	TCP	1434 [TCP segment of a reassembled PDU]
192.168.200.58	204.79.197.210	TCP	1434 [TCP segment of a reassembled PDU]
192.168.200.58	204.79.197.210	TCP	1434 [TCP segment of a reassembled PDU]
192.168.200.58	204.79.197.210	TCP	1434 [TCP segment of a reassembled PDU]
192.168.200.58	204.79.197.210	TCP	1434 [TCP segment of a reassembled PDU]
192.168.200.58	204.79.197.210	TCP	1434 [TCP segment of a reassembled PDU]
192.168.200.58	204.79.197.210	TLSv1.2	1434 Application Data
192.168.200.58	204.79.197.210	TCP	1434 [TCP segment of a reassembled PDU]
192.168.200.58	204.79.197.210	TCP	1434 [TCP segment of a reassembled PDU]
192.168.200.58	204.79.197.210	TCP	1434 [TCP segment of a reassembled PDU]
192.168.200.58	204.79.197.210	TCP	411 [TCP segment of a reassembled PDU]
192.168.200.58	204.79.197.210	TCP	1434 [TCP segment of a reassembled PDU]

Figure 3. File Attach in Hotmail

2) UDP Protocol Signature

This type of signature can be observed when a user makes an audio/video call or shares file within the Skype application, a client to client type of connection is established between the both clients. The 'size of packets' is the factor that is used for distinguishing the various activities. The size of the packets is between 1,165 to1,365 bytes during a video call and 129 to147 bytes during an audio call. The activity, 'sending a file' has maximum MTUs and it is between the two Skype users directly as shown in Fig. 4.



UDP	1412 Source port: 12354 Destination port: 50196
UDP	1412 Source port: 12354 Destination port: 50196
UDP	1412 Source port: 12354 Destination port: 50196
UDP	1412 Source port: 12354 Destination port: 50196
UDP	1412 Source port: 12354 Destination port: 50196
UDP	1412 Source port: 12354 Destination port: 50196
UDP	1412 Source port: 12354 Destination port: 50196
UDP	1412 Source port: 12354 Destination port: 50196
UDP	1412 Source port: 12354 Destination port: 50196
UDP	1412 Source port: 12354 Destination port: 50196
UDP	1412 Source port: 12354 Destination port: 50196
UDP	1412 Source port: 12354 Destination port: 50196

Figure 4. Sending file in Skype

In our previous paper [30], it was found that creating interactions was possible; but no further investigation was conducted in measuring these interactions and finding if they could be used to profile individuals at that time. Indeed, the user interaction has consisted of different features that represent the whole user action between the sender and receiver as shown in the Table I below. This area of investigation forms and the level of information uniqueness are the fundamental part of this paper.

TABLE I. USER INTERACTION FEATURES

No.	Feature Name	Example
1	Start Time of interaction.	2014.11.11.10:48:19.769086
2	End Time of interaction.	2014.11.11.10:48:19.817979
3	Source port number.	58823
4	Service IP.	216.58.208. %
5	Service port number.	443
6	# packets send from (client to server)	2
7	Total size of packets sends (client to server)	2000
8	# packets send from (server to client)	6
9	Total size of packets sends (server to client)	2868

IV. VALIDATION OF USER INTERACTIONS

The objective of this experiment is to validate the use of user based interactions as a feature-set and using it within a behavioural profiling biometric system in order to identify the users.

A. Dataset

The effectiveness and robustness of the experiment is largely dependent upon the quality and quantity of data. As such, a significant data collection activity was undertaken. After a search for open-access datasets failed to identify any suitable sources, the data was collected centrally from authors research centre. This enabled the researchers to set static IPs within the network in order to provide the ground truth and avoid changing IPs. Twenty-seven participants took part in the collection for 2 months during November 2014 to January 2015. This process was focussed purely on the collection of network metadata. The size of the complete dataset attained at the end was 62.4 GB.

Amongst this traffic, the experiment sought to focus upon nine services that were previously analysed for user interaction signatures; after excluding eBay application due to high level of noisy within identifying signature phase; which include user actions activities in each application from start to finish, source and destination ports, application IP and number and size of packets in each direction as shown in Table I. Table II illustrates a breakdown of the data collected against the chosen nine applications. YouTube, Dropbox, Facebook and Google applications have the largest volume of traffic over than, 21, 17, 5 and 1.8 million packets respectively. After applying the user interaction signatures that were derived from the previous section a vast reduction can be seen as illustrated in the column 3 of Table II. This highlights a reduction in data processing in comparison to packet-based approaches and also highlights the focus upon interactions rather than flows, which is similar to flow-based approaches. As the data collection was based upon the capture of real data, not all participants were active across all the nine services. However, it is clear that there is sufficient data to test the experimental hypothesis.

TABLE II. STATISTICAL INFORMATION FOR EACH APPLICATION

Application	No. of packets	No. of Interaction	No. of Participants
YouTube	21,131,316	1,322,848	27
Facebook	5,727,953	386,741	27
Google	1,857,420	194,404	27
Twitter	747,584	71,403	27
Wikipedia	1,250,302	5,719	20
Hotmail	703,711	122,986	19
Dropbox	17,480,739	98,555	16
BBC	201,263	4,180	12
Skype	575,030	178,686	12

B. Classification Phase

The classifier has to be applied on the organised data by eliminating the redundant data and converting all input to numeric and this has been done by normalisation phase before the classifier starts [31]. Neural network classifier is considered as an effective approach that deals with complicated patterns. Feed-forward multi-layered perceptron neural network (FF_MLP) was selected for the study because it is useful in complex information processing through using a several layers of adaptive weights which could solve complex non-linear problems [32-33]. A one layered neural network was built with nine inputs and one output. Within the FF ML network, supervised learning technique called Levenberg Marquardt (trainlm) was used for training the network and to solve non-linear problem. Neuron sizes have taken the follow values10, 15, 20, 25 and 30. Table III defines the different neural network parameters with varying numbers of neurons utilised to allow for optimisation of the problem.

Table III. Classifier Setting

Classifier Name.	No. Neuron	No. Layer	Activation Function	Training Function	Epochs
FF-MLP	10-30	1	tansig	trainlm	1000 Iterations

C. Methodology

The methodology utilised to validate the level of user interactions uniqueness is based upon the standard approach for biometric testing which gives the system an ability to reject imposters and match with the authorised users, which is initiated for all the participants [34]. The experiment is repeated for each participant, and in every main participant plays the role of the authorised user and the remaining acts as the impostors. Thus, dividing each user interactions in each application to two halves; one for training the classifier and the other for testing; is a core step because training and testing samples should be separated to ensure that there is no single sample has utilised in both parts. Also, in order to ensure there is sufficient data for both training and testing, a minimum of 30 user interactions is set as a threshold: users will be excluded on an application test if they had less than 30 interactions for that application. The evaluation phase of the classifier performance utilises the True Positive Identification Rate (TPIR) - if the biometric system outputs the identities of the top t matches for each user sample, where t is representing the rank of accuracy. The following sections are explained each step in details.

V. EXPERIMENTAL RESULTS

There are a number of key results from our experiment. Table IV illustrates the TPIR for all users in each application. Applications that have scored a high performance in rank1 considered as a high accurate application in terms of user profile because this reflects the level of discriminative information of user which contribute the classifier towards correct choice. Skype and Hotmail applications have scored high TPIR from rank1 with 98.1% and 96.2% TPIR respectively which means the system has correctly classified almost all their samples by assigning them to the correct user. Although, BBC application has got good level of accuracy 81.8% in rank1, this value has gradually increased to 88.7%, in rank2 and continuous improved up to 95.4% in rank5. There are some applications scored low TPIR such as Google and Wikipedia where the system have correctly identified more than 66% of their samples from the first top value (rank1) but this proportion is considered as a promising value if we know that they have utilised by the majority of the participants 27 and 20 respectively as show in Table I and the level of improvement that scored in their accuracy in ranks 2 and 3. Although, YouTube application have got the lowest level of accuracy where its TPIR is 62.8% in rank1, its interactions represent more than 55% of whole users' applications as illustrated in Table II.

While some applications results have got enough discriminative information to be correctly identified in ranks 1 and 2 such as Skype and Hotmail, the remains applications almost have scored significant improvement in their performance from rank1 to 5 between 2% to 25%, for instance BBC, Wikipedia, Twitter, YouTube and Drobox as shown in TABLE IV.

TABLE IV. USERS IDENTIFICATION RATE IN EACH APPLICATION

App. Name	No. Users	Rank1 %	Rank2 %	Rank3 %	Rank4 %	Rank5 %
Skype	12	98.1	98.2	98.2	98.2	98.2
Hotmail	19	96.2	96.8	96.9	96.9	97
BBC	12	81.8	88.7	92.5	93.7	95.4
Google	27	71.7	77	79.4	80.7	82.2
Wikipedia	20	66.9	78.2	83.6	86.6	89.2
Facebook	27	66.7	69.6	70.8	71.3	71.9
Twitter	27	65.3	75.3	79.5	81.9	83.4
YouTube	27	62.8	71.4	74.8	77	78.9
Dropbox	16	57.1	67.5	73.9	78.8	82.8

For individual service, there is a numbers of promising results have scored across all applications. Indeed, all participants have experience at least three applications. Table V shows the top three applications that scored high level of accuracy from rank1 value for each user. The results reveal that 1/3 of participants have been correctly identified via their interactions with level of accuracy more than 80% in all top three applications. It is also shown that more than 75% of the users have got at least one application with 80% TPIR. While the system has an ability to profile 92% of the users from their interactions with TPIR more that 74%, there is a small proportion where the system can only assign less than 60% of the interactions to the correct participant.

TABLE V. USERS TPIR IN RANK 1 TOP THREE APPLICATIONS

User ID	First App.		Second App.		Third App.	
	Name	TPIR%	Name	TPIR%	Name	TPIR%
1	Wikipedia	100	Google	95.2	YouTube	50
2	Skype	94.1	BBC	74.6	Wikipedia	74.1
3	Skype	100	Hotmail	60.8	Google	59.1
4	Hotmail	91.8	Google	91.6	YouTube	90.9
5	YouTube	74.1	Google	73.6	Hotmail	13
6	Skype	100	Google	89.1	Hotmail	85.5
7	Google	79.2	BBC	64.8	Wikipedia	50
8	Wikipedia	100	Facebook	79.5	YouTube	56.5
9	Skype	100	Hotmail	95	Wikipedia	93
10	Hotmail	83	Skype	63.7	Google	56.1
11	Hotmail	80.5	Skype	80.3	Google	72.7
12	Skype	99.7	BBC	95	Hotmail	80.2
13	Dropbox	80.9	Google	75.3	Facebook	72.7
14	Skype	100	Hotmail	62.3	Dropbox	48.2
15	Facebook	74.5	YouTube	71.1	Dropbox	70.9
16	Wikipedia	95.6	Hotmail	35.4	YouTube	28
17	Google	59	Wikipedia	52	YouTube	30.6
18	Wikipedia	98	BBC	82	YouTube	67.2
19	Skype	99.4	Hotmail	95	BBC	61.7
20	Dropbox	75.2	Facebook	73.9	Google	63.7
21	Skype	100	Hotmail	85.3	Twitter	79.3
22	Twitter	43	Google	29.4	YouTube	4.2
23	Facebook	75.5	Hotmail	71.4	Twitter	58.6
24	Hotmail	100	Skype	100	BBC	91.8
25	Google	80.8	Dropbox	75.9	YouTube	51.1
26	Wikipedia	76.1	Google	64.8	Facebook	62.4
27	Skype	100	Google	100	YouTube	100

In the level of individual user, the average TPIR in first application for all users has scored 87.4%, this result has proved that it is possible to use user interaction for creating user profile which is the main aim of this work. Therefore, there is some participants who scored high level of TPIR across the three applications from rank1as shown in (Table V) such as users 4, 9, 24 and 27, but user 27 has got the best user profile because of the highest level of accuracy he got it in all top three applications when the system was able to completely identified all his interactions with 100% accuracy.

Although the majority of participants has scored a promising result in the first application, Skype application has scored the best user profile result among the 9 application with almost over 90% TPIR and this is due to the nature and precise of it signature also the port connection is often constant with particular number above of 1024 [35].

VI. DISCUSSION

The experiment reveals that the natural of the user interaction derived from application level is unique, thereby using it to build a user behavioural profile is a promising solution to identify the insider misuse. Moreover, the experiment evidently shown that by using user interaction the system was able to identify some participants in different applications with 100% level of accuracy as shown in Table V. Since this high level of accuracy achieved form rank1, it is a clear indication that there is a discriminative information exist in the user interaction which contribute towards right classification. The explanation of TPIR being high is attributed to the level of information uniqueness of particular user in particular application. Therefore, obtaining a precise user actions pattern lead to the accurate user profile as demonstrate in Table V.

This provides forensic investigators with a strong approach to identify relevant traffic. When combined with the IP address and windowing (an approach that uses the successful authenticated user interaction to identify the IP address and then uses a windows +/- period (1-2 minutes) to tag all traffic from that IP) provides a very successful approach to target upon the traffic that is most relevant. The use of the IP address in this context is viable as the assumption will only help for a short period of time rather than for the complete duration of the network capture.

This approach of using both the biometric and the IP address suggests that it is not necessary to correctly classify every user interaction but it is important to have at least one application with enough discriminative information for the range of users. Indeed, this approach has providing some positive insights in terms of creating user-behavioral profile from metadata while the system successfully identifying some individual with 100% of accuracy. Subsequently, reducing the numbers of network traffic analyzing based on user behavioral profile from application level interactions by using a metadata without relying on IP address is a good objective in terms of millions of records needs to be investigated. Therefore, merely sufficient along the timeline for the IP address windows to overlap and provide a confirmation of the IP address.

VII. CONCLUSIONS AND FUTURE WORK

This research has investigated the value of the features that can be derived from raw data based upon user interactions. Accordingly, the outcomes of this experiment identify the need to further explore the discriminative values of features and classifier optimisation strategies, such as k-Nearest Neighbour algorithm, Support Vector Machine and Ensemble classifiers.

Further research will also be undertaken in understanding the nature of the user interactions that exist on different technology platforms (i.e. smartphone, tablet, laptop) to identify whether it is possible utilise a single biometric profile, or would it be technology dependent.

REFERENCES

[1] Internetlivestats, “Internet live stats”, [online], http://www.internetlivestats.com/internet-users/, date accesed 2nd March 2016.

[2] Microsoft, “Office 365 for business FAQ,[online], https://products.office.com/en-us/business/microsoft-office-365-frequently-asked-questions, date accessed 1st March 2016.

[3] Library, “Introduction to Google Docs”,[online], http://www.lfpl.org/jobshop/docs/google-docs.pdf, date accessed 1st march 2016.

[4] National Statistical in the UK, [online] , http://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/bulletins/internetaccesshouseholdsandindividuals/2015-08-06, date accessed 15th March 2016.

[5] Ofcom (2015): The UK is now a smartphone society,[online], http://media.ofcom.org.uk/news/2015/cmr-uk-2015/, date accessed 20th March 2016.

[6] Statistical portal,2016, “Application downloaded”,[online], http://www.statista.com/statistics/241587/number-of-free-mobile-app-downloads-worldwide/, date accessed 25th March 2016.

[7] Microsoft 2016, “ Security Threats”,[online] https://msdn.microsoft.com/en-us/library/cc723507.aspx, accessed on 14th April 2016.

[8] Mcafee, "Net Losses:Estimating the Global cost of cybercrime", [online] http://www.mcafee.com/us/resources/reports/rp-economic-impact-cybercrime2.pdf, date accessed: 5 March 2016

[9] Verizon, “2014 Data Breach Investigations Report”, [online], http://www.verizonenterprise.com/DBIR/2014/, date access 28th March 2016.

[10] PwC, “2015 Information security breaches survey”, [online], https://www.pwc.co.uk/assets/pdf/2015-isbs-technical-report-blue-03.pdf, date accessed: 26th March 2016.

[11] L.D. Merkle, “Automated Network Forensics”, Proceedings of the conference on genetic and evolutionary computation (GECCO 2008), pp 1929-1932, in press.

[12] K. Wang and S. J. Stolfo, “Anomalous Payload-Based Network Intrusion Detection”, Proceeding of the 7th International Symposium RAID 2004, p.p. 203-222, Sophia Antipolis, France, September 15 - 17, 2004, in press.

[13] QOSMOS (2015) Security information and Event managemnt (SIEM) use case, [online on] http://www.qosmos.com/wp-content/uploads/2015/08/Qosmos_SIEM_Use-Case_2015.pdf, date accessed 26th March 2016.

[14] Cisco, “Cisco Visual Networking Index: Forecast and Methodology, 2013-2018”, [online], http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.html, date accessed: 20 December 2014

[15] G. G. Baehr, W. Danielson, T. L. Lyon, G. Mulligan, M. Patterson, G. C. Scott and C. Turbyfill, “System for packet filtering of data packet at a computer network interface”, patent: US5884025, 1995

[16] K.C. Claffy, H.W. Braun and G.C. Polyzos, "A parameterizable methodology for Internet traffic flow profiling", Selected Areas in Communications, IEEE Journal on , vol.13, no.8, pp.1481-1494, Oct 1995.

[17] Awodele, O., Oluwabukola, O., Ogbonna, A. C., & Adebowale, A., “Packet Sniffer – A comparative characteristic evaluation study”. Proceedings of Informing Science & IT Education Conference (InSITE) 2015, 91-100.

[18] INFOSEC (2016): Password Cracking Using Cain & Abel, [online], http://resources.infosecinstitute.com/password-cracking-using-cain-abel/,date accessed 20th March 2016

[19] I. Ahmed and K. Lhee, “Classification of packet contents for malware detection”, Journal in Computer Virology, NOVEMBER 2011, VOLUME 7, ISSUE 4, PP 279-295

[20] S. Dharmapurikar, P. Krishnamurthy, T. Sproull and J. Lockwood, “Deep packet inspection using parallel Bloom filters”, Proceedings of 11th Symposium on High Performance Interconnects, pp.44,51, 20-22 Aug. 2003, in press.

[21] F. Yu, Z. Chen, Y. Diao, T. V. Lakshman and R. H. Katz, “Fast and memory-efficient regular expression matching for deep packet inspection”, ancs '06 proceedings of the 2006 acm/ieee symposium on architecture for networking and communications systems, pp 93-102, in press

[22] Y. H. Cho, S. Navab and W. H. Mangione-Smith, “specialized hardware for deep network packet filtering", proceeding of the 12th international conference, pp 452-461, montpellier, france, september 2–4, 2002, in press.

[23] B. Claise, “Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of IP Traffic Flow Information”, IETF, RCF5101, January 2008, [Online], https://tools.ietf.org/html/rfc5101, date accessed: 17 January 2015

[24] Cisco "NetFlow version 9", [online], http://www.cisco.com/c/en/us/products/ios-nx-os-software/netflow-version-9/index.html, date accessed 25 February 2015

[25] Sflow, “sflow”, [online] www.sflow.org, date accessed 24 February 2015

[26] Juniper, “Juniper flow monitoring”, [online] http://www.juniper.net/us/en/local/pdf/app-notes/3500204-en.pdf, date accessed 25 February 2015

[27] B. Li, J. Springer, G. Bebis and M. H. Gunes, “A survey of network flow applications”, Journal of Network and Compouter Applications Vol.36 issue 2 pp. 567-581.

[28] eBizMBA(2016): Top 15 Most Popular websites, [online] http://www.ebizmba.com/articles/most-popular-websites,date accessed 1st September 2016.

[29] Alotibi G, Li F, Clarke NL, Furnell SM (2015): “Behavioral-Based Feature Abstraction from Network Traffic”, 10th International Conference on Cyber warfare and Security, Kruger National Park, South Africa, 24-25 March, pp1-9, ISBN 978-1-910309-97-1, 2015

[30] Li.F, Clarke.N, Alotibi.G & Joy.D(2015): “Forensic Investigation of Network Traffic: A Study into the Derivation of Application-Level features from Network-Level Metadata”, Big data, Could and Security (ICT-BDCS 2015), 27-28 July, ISSN: 2382-5669, pp68-73, 2015.

[31] J. Sola and J. Sevilla (1997) “Importance of Input Data Normalization for the Application of Neural Networks to Complex Industrial Problems”, IEEE TRANSACTIONS ON NUCLEAR SCIENCE, VOL. 44, NO. 3, JUNE 1997.

[32] Y. Chtioui, D. Bertrand, M. Devaux, D. Barba (1997): Comparison of multilayer perceptron and probabilistic neural networks in artificial vision application to the discrimination of seeds, Journal of Chemometrics 11 (1997) 111 – 129.

[33] Amit Mehta, Arjun Singh Parihar and Neeraj Mehta (2015): Supervised Classification of Dermoscopic Images using Optimized Fuzzy Clustering based Multi-Layer Feed-Forward Neural Network, IEEE International Conference on Computer, Communication and Control, Indore, 10-12 Sept. 2015.

[34] Nathan Clarke, Steven Furnell & Benn Lines (2004) “ Application of Keystrock Analysis to Mobile Text Messageing”, Proceedings of the 3rd Security Conference, Las Vegas, USA, 14-15 April, 2004.

[35] Skype, “Which ports need to be open to use Skype for Windows desktop”, [online] https://support.skype.com/en/faq/FA148/which-ports-need-to-be-open-to-use-skype-for-windows-desktop, date accessed 4th March 2016