https://doi.org/10.17576/apjitm-2022-1102-08
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 11 No. 2 December 2022: 97 - 111
e-ISSN: 2289-2192

Received: 22 June 2022
Accepted: 14 October 2022
Published: 1 December 2022

VALIDATION OF INDIVIDUAL IDENTIFICATION THROUGH DECISION TREE PACKET HEADER PROFILING

KHAIRUL OSMAN
T'NG QI FENG
HAIREE IZZAM MOHD NOOR
NOOR HAZFALINDA HAMZAH
GINA FRANCESCA GABRIEL

ABSTRACT

The drastic rise in the cybercrime rate associated with the surge of users' dependence on the Internet has elevated the concern of digital forensic examiners toward the footprints of perpetrators left in a virtual environment. However, suspect identification is a big challenge in network forensics due to the anonymous nature of data transmission across the network. This study utilises the decision tree classification approach to characterise users from their behavioural web navigation pattern using the meta-data of captured network packets (Destination IP, Protocol, Port Source, and Port Destination). A total of 95,795,379 network packet headers from 96 subjects were successfully collected. Their meta-data header packets were statistically profiled to generate digital fingerprints that try to link their action on the network to their identity accurately. Hence, CHAID decision tree modelling using Destination IP, Unique protocols, and a combination of the two, including Port source and Port destination, resulted in an accuracy of 4.07%, 6.34%, and 6.36%, respectively. However, the modelling could not create a reliable decision tree for the Port source and destination. The validation study on all the combined variables had a similar accuracy of 6.36%, indicating model created had reproducibility capability. Despite the outcome, the proposed method is not yet sufficiently strong for suspect identification. Further enhancement to improve its accuracy is required.

Keywords: digital forensic, decision tree, digital fingerprint, user identification

INTRODUCTION

In recent years, internet usage has drastically increased to the extent that it has become the primary medium for our daily communication and commerce activities. The statistic from Internet World Stat reveals the growth of Internet users from about 16 million in 1995 to 5,168 million in 2021 (Brahimi, 2022). In addition, the COVID-19 outbreak also expedited the exponential growth of the Internet as it fostered a new normal by transforming many physical tasks and activities into online mode. In Malaysia, MCO (Movement Control Order) has caused Internet users to spend more hours online in 2020 compared to 2018. This upward trend has corresponded with the rise in the percentage of regular users (spent 5-12 hours daily) and heavy users (spent >12 hours daily) from 37% and 14% to 50% and 21%, respectively (Malaysian Communications & Commission, 2020).

As the growth of the Internet encouraged the digital transformation of our daily routine, it also afforded vast opportunities for cybercrime. A study in the UK revealed that cybercrime cases recorded a remarkable incline during COVID-19 due to the shift of crime opportunities from physical to online (Buil-Gil et al., 2021). In Malaysia, the number of cybercrime cases showed an upward trend from 11,875 in 2019 to 14,229 in 2020, and this trend continued in the next year, with 4,327 cases reported in the first quarter of the year alone (Dawn Chan, 2021).

Identification of the perpetrator is always the challenge of digital crime due to its borderless nature and the anonymity of the digital data (Caviglione et al., 2017). However, their malicious action on the Internet will leave traces in a digital form. Searching evidence from the history, cookies, cache, and download list from web browsing activities has become a crucial component of digital forensic investigation (Mugisha, 2019). In 2020, two Malaysian were charged by the United States Department of Justice (DoJ) for hacking video game companies in the United States, France, Japan, Singapore, and South Korea to obtain the game resources illegally (Muzliza Mustafa, 2020). This case shows that even cybercriminals who conduct intrusions across the national border can be detected and identified with sufficient digital traces.

Data transmission through the Internet depends on data communication protocols and network applications. The WWW (World Wide Web) transfers website resources hypertext for user access (Nath, 2015). When dealing with FTP (File Transfer Protocol), data transmission involves the direct transfer of computer files from one host to another based on a client-server model (Rahim et al., 2018). In VPN (Virtual Private Network), the encrypted data is transmitted through a virtual point-to-point connection established by tunnelling the network traffic (Abdulazeez et al., 2020). According to the Global Internet Phenomena Report in 2019, WWW is the second most prevalent network application, constituting 13.1% of downstream traffic and 10.3% of upstream traffic. (Sandvine, 2019).

When the user is browsing the WWW, the data transmitted across the network is divided into small segments called packets. The packets typically consist of the payload and the header. Payload refers to the intended information transmitted, while the packet header is a packet label that provides information about the packet's content, source, and destination. Appropriate capturing and analysis of the network packets can provide valuable and relevant information to link a suspect with his criminal act during the forensic investigation (Sikos, 2020).

Network forensic analysis usually focuses on the packet header before the payload because the conventional approach is ineffective in analysing the load (Cha & Kim, 2017). IPsec (Internet Protocol Security) typically encrypts the payload in its transport mode or TLS (Transport Layer Security) to secure the data communication (Varadhan, 2016). Conversely, the packet headers are generally not encrypted except when tunnelled through a VPN. Therefore, the meta-data of the packets, such as the Source and Destination IP address, port numbers, and protocol types, can be easily extracted for analysis.

The unique web browsing pattern of the Internet user has the potential to create a digital fingerprint that can identify them from their future web browsing activities through several classification algorithms (Santise et al., 2012). As the information in the packet headers can reflect the users' web browsing activities, this research proposed that the profiling of the packet headers can generate a robust model viable for user identification. The statistical classification algorithms widely used for the network packet include Naive Bayes (Cha & Kim, 2017; Fadlil et al., 2017; Meti et al., 2017), Support Vector Machine (Bakopoulou et al., 2019; Cha & Kim, 2017), k-Nearest Neighbour, and Decision tree (Cha & Kim, 2017; Cheng & Wang, 2015; Kathuria & Gambhir, 2016). Among various algorithms, the decision tree is the most favoured approach because of its superior classification performance enhanced by the efficient tree traversal algorithm and the ease of interpretation due to the intuitive rule in its node splitting. On the other hand, the decision tree is prone to outliers and difficult in size control when dealing with complicated datasets, but this issue can be minimised by conducting data clean-up before analysis (Ray, 2019). Hence, this research aims to generate a digital fingerprint based on the users' web browsing behaviours using decision tree analysis and determine its ability to identify an individual by analysing their network packet meta-data.

METHODOLOGY

A quantitative approach was used to construct the unique digital fingerprint of selected individuals through packet header profiling of web browsing activity. The methodology is summarised in Figure 1. Generally, the study protocol comprises several steps encompassing data collection, merging, processing, analysis, construction of individual digital fingerprint models, re-testing the accuracy of models, and validation (Figure 1).

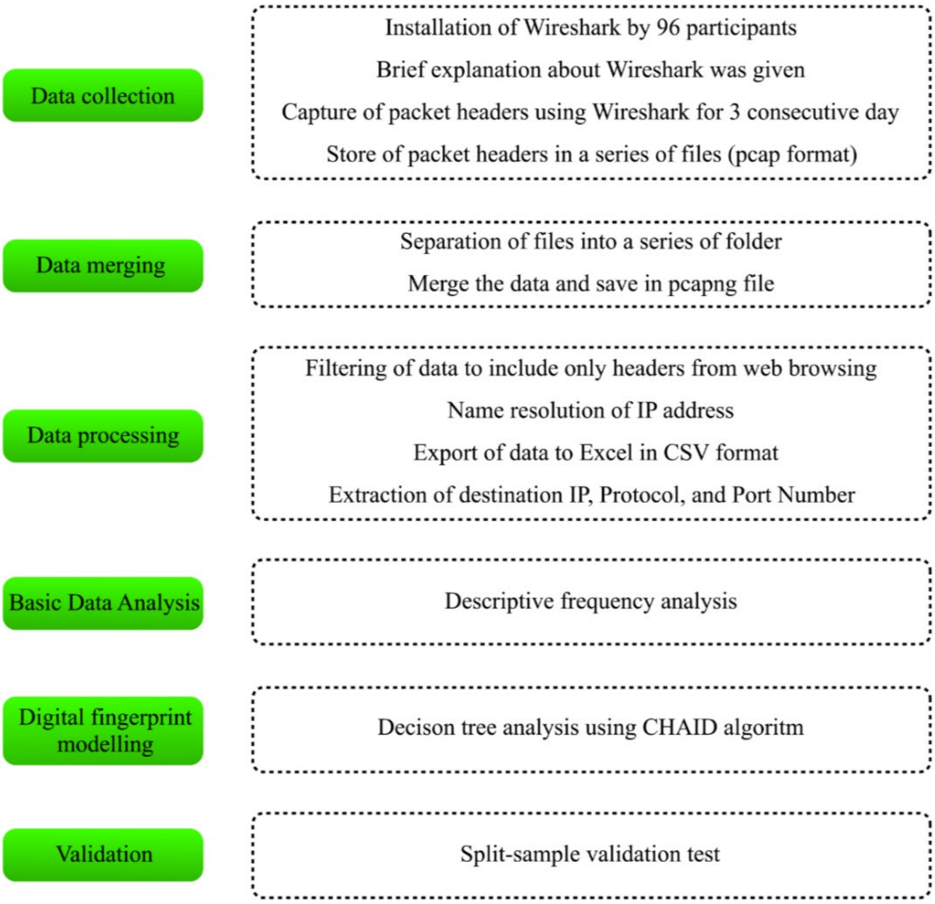
Data collection
Installation of Wireshark by 96 participants
Brief explanation about Wireshark was given
Capture of packet headers using Wireshark for 3 consecutive day
Store of packet headers in a series of files (pcap format)
Data merging
Separation of files into a series of folder
Merge the data and save in pcapng file
Data processing
Filtering of data to include only headers from web browsing
Name resolution of IP address
Export of data to Excel in CSV format
Extraction of destination IP, Protocol, and Port Number
Basic Data Analysis
Descriptive frequency analysis
Digital fingerprint modelling
Decison tree analysis using CHAID algoritm
Validation
Split-sample validation test

FIGURE 1. Summary of Methodology

DATA COLLECTION

Network packet headers were collected from 96 participants consisting of Universiti Kebangsaan Malaysia (UKM) undergraduates. The sampling method was convenient sampling, where the participants were chosen based on their availability and willingness to participate. As the capture of network packets involved privacy issues, all the participants were requested to sign a consent form before participating in this research.

The capture of packet headers was performed using software called Wireshark. The software was installed on each participant's computer under the guidance of the researchers. A brief explanation of the operation of Wireshark was given to each participant. The software was then required to run in the background for three consecutive days. Filter to the captured data packet was applied to ensure only header data was extracted from the packets and later stored as pcap formatted files. The body of the packet was not stored. The selected parts of the packets were captured using the Wireshark capture options, putting a checkmark next to "Limit each packet to" and setting the value at 54 bytes.

DATA MERGING

As the data size in each file was quite large, merging data requires that the files be separated into a series of folders. The data were combined using the Mergecap software included in the Wireshark package. The merged files were then saved as pcapng formatted files.

DATA PROCESSING

The collated data in pcapng was then loaded into Wireshark. TCP filter was applied to only include headers from web browsing activity. This action allowed us to capture packet headers for TCP (Transmission Control Protocol), TLSv1.2 (Transport Layer Security version 1.2), and TLSv1.3 (Transport Layer Security version 1.3).

Name resolution of the IP address was then resolved by re-analysing the captured DNS (Domain Name System) packets and utilising an external network name resolver. The external network name resolver was Google DNS (8.8.8.8 and 8.8.4.4). The data were then exported as a CSV formatted file.

The CSV (Comma Separated Values) files were then loaded into Microsoft Excel. Then the information in the packet headers was extracted to obtain data on Destination IP, Protocol, Port source, and Port number. Extracted data were then piped into SPSS for subsequent analysis. As Excel can only handle approximately 1,048,576 rows of data, the data Excel were split into different files and later combined inside SPSS.

BASIC DATA ANALYSIS

Descriptive frequency analysis was carried out to provide an overview of the web browsing use of data collected as a whole and from each participant. The study included max total packets collected, the most frequented website and the frequency of visits based on IP.

DIGITAL FINGERPRINT MODELLING

Decision tree analysis was conducted on the Destination IP, Protocol, Port source, and Port number data to construct a reliable digital fingerprint model. The previous mentioned independent variable was used individually or in combination to create multiple variations of the fingerprint model. The algorithm used to create the digital fingerprint as a decision tree was CHAID (Chi-square automatic interaction detection). As mentioned previously, the independent variables (predictor) were the Destination IP, Protocol, Port source, and Port destination, while the dependent variable was the user.

Each decision tree analysis generated a digital fingerprint model that can predict the users based on the distribution pattern of their packet headers. The accuracy of the digital fingerprint was viewed as the percentage of predicted users correctly matched with the observed one.

VALIDATION

The split-sample validation test examined the validity of the digital fingerprint model. A split-sample validation test was made by randomly partitioning the data into 80% training sample and 20% test sample. The selection of samples of either training or test sample was done automatically using SPSS. The model was generated using a training sample and tested on the test sample. The split-sample validation test provided an unbiased estimation of the model performance for actual prediction in real-life scenarios (Vabalas et al., 2019). The validity of the protocol to create the various model was determined based on the consistency of the model accuracy across the training and test sample.

RESULT AND DISCUSSION

The results were based on 96 users' web browsing activity (Table 1). Subject demography consisted of 29 Malay (30%), 54 Chinese (56%), and 14 India (14%). For gender, there are 55 (57%) female users, more than male users (41; 43%). Furthermore, all users have an identical highest educational background in which they are all undergoing various degree programs. As all users are degree students, they mainly come from 22-24 years old with a frequency of 89 (94%), whereas the remaining 6 (6%) users are between 19- to 21 years old. The use of subjects between these age groups is preferred, as reflected by (Clarke et al., 2017) research on packet meta-analysis and (Vinupaul et al., 2017) network flow analysis. These studies emphasised the importance of demography and its correlation with the prevalence of cyber criminals in a selected population (Adeniyi, 2019).

TABLE 1. Demography for (a) Gender, (b) Race, (c) Age, and (d) Educational Background of 96 users

Criteria	Sub-criteria	Detail
Gender	Male	43%
	Female	57%
Race	Malay	30%
	Chinese	56%
	India	14%
Age	19 – 21	6%
	22 – 24	94%
Educational level	Degree	100%

There was a total of 95,795,379 packet headers collected. Among the users, User ID 34 acquired the most significant number of packets which is 3,897,463, making up 4.07 % of the total packets meta-analysed for this study. User ID-5 is the participant with the smallest number of packet headers collected, constituting only 1,274 packet headers (1.33 x 10^{-3}%). The distribution of the packet headers has an SD of 886,163, inferring that the number of packet headers collected by each user is highly spread out over a great range of values. We must clarify that this observation contrasted with past studies (Ikuesan et al., 2020; Malatras et al., 2017) that have uniform packet distributions among the users. The contradicting result was probably due to an additional requirement added to our study, which did not limit nor specify the duration of a user to use the Internet and the minimum number of packets needed to contribute to the study. In other words, all users could freely browse the web according to their interests. It enabled the inclusion of the users' varying browsing frequencies as part of their habits to enlarge the information entropy, allowing better chances to create a reliable digital fingerprint (Laperdrix et al., 2016).

TABLE 2. Top Three Most Frequently Visited Destination IP

Destination IP	Potential Website/server	Number of Packets	Percentage (%)
10.33.41.193	Private Network	2,681,558	2.80
192.168.1.16	Private Network	2,414,170	2.52
10.33.42.202	Private Network	1,990,778	2.08

TABLE 3. Unique Destination IP

Unique Destination IP	Potential website / server	Corresponded user	Number of packets
a184-29-99-86.deploy.static.akamaitechnologies.com	CDN service from akamai.com	User 62	771,493
mirror.karneval.cz	Mirror site for Kali Linux distribution	User 42	725,084
2401:3c00:c:b690:25f9:71c8:5516:9a91	Broadband service from Webe Digital Sdn. Bhd.	User 48	691,761

As shown in Table 2, the top three most frequently visited Destination IPs are all private IP addresses commonly used for residential and corporate internal networks. 10.33.41.193 and 10.33.42.202 are from the Class A private IP range (10.0.0.0 – 10.255.255.255) with common usage by routers of large organisations that contain many connected devices. 192.168.1.16 is the Class C private IP range (192.168.0.0 – 192.168.255.255) commonly used by the domestic router. These are likely routers used in houses or low-rise buildings. The prevalence of these three private IPs is probably due to the high frequency of internal communications between the users' devices and the web servers during their web browsing activities.

Each user would have unique IPs that are only present in these individual acquired packets, and these unique IPs can enhance the discrimination of his identity from other users. Table 3 lists some of the unique Destination IPs (only the public IP with the top three packets). Among them, a184-29-99-86.deploy.static.akamaitechnologies.com is the CDN (Content Delivery Network) service from Akamai Technologies, Inc. User ID 62 was likely to frequently access one of the cache servers of Akamai CDN for reduction of network latency and faster web browsing experience (Zolfaghari et al., 2020). The mirror.karneval.cz visited by User ID 42 is a publicly accessible mirror site for Kali Linux distribution. He probably had downloaded Kali Linux packages from this mirror site to update and install Kali Linux OS on his main device. 2401:3c00:c:b690:25f9:71c8:5516:9a91 is possibly the server of Webe Digital Sdn. Bhd. that provides broadband service to User ID 48. According to the TM's corporate report on 11th September 2017, the broadband service of Webe had a 5.6% household penetration rate. As Webe rebranded its service to UniFi afterwards, its household penetration is expected to be lower than previously stated. Therefore, it is not surprising that only 1 out of 96 users is using the broadband service of Webe.

TABLE 4. Decision Tree Analysis Using Destination IP as Independent Variable

	Predicted			
Observed	User 1-33	User 34	User 35-96	Accuracy (%)
User 1-33	0	30,435,700	0	0.00
User 34	0	3,897,463	0	100.00
User 35-96	0	61,462,216	0	0.00
Overall percentage (%)	0.00	100.00	0.00	4.07

A CHAID decision tree model (Table 4) using Destination IP as the independent variable and user as the dependent variable was created. This model was developed to test the model's capability of identifying a user. During the tree-growing process, one of its stopping criteria that limited the number of categories terminated the tree-splitting to avoid building a complex and unreliable tree. As a result, the Destination IP was excluded from the tree-splitting and resulted in only a single node mapped to User ID 34. The accuracy obtained from this model is only 4.07 %. s This structure is solely contributed by User ID 34, which achieved an accuracy of 100% for his identification.

TABLE 5. Top Three Most Frequently Used Protocols

Protocol	Number of packets	Percentage (%)
TCP	85,125,318	88.86
TLSv1.2	9,078,471	9.48
HTTP	1,162,565	1.21

TABLE 6. Unique Protocol

Unique Protocol	Corresponded user	Number of packets
DRBD	User 52	66,911
PKIX-CRL	User 47	8
SSLv3	User 7	3

Based on Table 5, protocol mainly consists of TCP (88.86%), TLSv1.2 (9.48%), HTTP (1.21%), and a trace amount of other uncommon network protocols. As TCP has enormous usage for the ordered and reliable data transmission between the Internet application and the Internet Protocol, it is the most frequently used protocol (Alotaibi et al., 2017). Meanwhile, TLSv1.2 provides secure encryption of the TCP traffic to enhance the privacy and confidentiality of web services such as electronic commerce and asset management (Stuart Jacobs, 2016). Therefore, the users of TLSv1.2 probably have habits of visiting e-commerce websites like Shopee, Lazada, and Amazon because online shopping has become popular since the covid-19 pandemic. During the covid-19 pandemic, the E-commerce markets of Malaysia had a tremendous growth rate of 37% compared to the pre-covid period due to the movement restriction and the improvement of online business infrastructure (Raj S. & Gohain, 2021). HTTP (Hypertext Transfer Protocol) is an application-layer protocol for transferring the hypermedia document of WWW during the interaction between the web server and the user (J. Chen & Cheng, 2016). About half of the users (56%) are Chinese, so we speculated their regular visits to some China HTTP websites such as baidu.com, 360.com, and xinhuanet.com as the main contribution to the HTTP ubiquity. In contrast with the current trend of HTTPS (Hypertext Transfer Protocol Secure) adoption, many China websites are still using HTTP due to the interception from the Great Firewall censorship and the lack of HTTPS support by the China popular browsers (P. Chen et al., 2014; Porter Felt et al., 2017).

Table 6 illustrates the unique protocols acquired by User ID 52, User ID 47, and User ID 7 in this research. Among them, only DRBD (Distributed Replicated Block Device) had a significant number of packets. This protocol synchronises data from an active primary node to a passive secondary node. The user of this protocol is to ensure data availability in the event of a failure in the primary node (Park et al., 2013). Its frequent usage is probably because User ID 52 had routine visits to websites that implemented DRBD for high data availability.

On the other hand, the frequency of PKIX-CRL (Public-Key Infrastructure X.509 - certificate revocation list) and SSLv3 (Secure Sockets Layer version 3) are low, indicating that they are not regularly used by User ID 47 and User ID 7. PKIX-CRL is a list of the revoked digital certificate for Internet PKI, possibly requested by User ID 47 to check the validity of a digital certificate. SSLv3 was broadly used for secure e-commerce service over HTTP before being succeeded by TLS, so we suspect that these packets were captured when User ID 7 browsed some old commerce websites.

TABLE 7. Decision Tree Analysis Using Protocol as Independent Variable

	Predicted (captured packets)									
User ID	7	34	37	41	42	47	52	3, 57, 72	Others	Accuracy (%)
7	71,929	259,199	31	7925	197,925	0	0	Trace amount	< 1	13.39
34	0	3,897,463	0	0	0	0	0			100.00
37	6	402,876	19,323	45,010	466,635	0	0			2.07
41	4	279,616	2	377,718	388,417	0	0			36.12
42	29	949,724	19	47,700	1,512,051	2	0			60.25
47	244	1,003,254	1361	14,629	238,695	125,385	0			9.06
52	4	664,141	1	160,311	422,570	0	66,911			5.09
Others	2603	77,669045	9239	509,272	5,852,197	122051	0			< 1.00
∑ %	0.08	88.86	0.03	1.21	9.48	0.26	0.07	< 1.00	0.00	0.08

The decision tree (Table 7) constructed using the protocol as the independent variable has an accuracy of 6.34%. User ID 34 achieved the highest accuracy (100%), followed by User ID 42 (60.25%), User ID 41(36.12%), User ID 7 (13.39%), User ID 47 (9.06%), User ID 52 (5.09%), and User ID 37 (2.07%). The other users either have a trace amount of correct predicted packets (<1%) or no correctly predicted packets. As User ID 34 did not have any unique protocol, the 100 % accuracy of his prediction is probably due to his extensive usage of TCP, which significantly outweighed the other users. User ID 42 and User ID 41 also achieved high accuracy, possibly a result of their exclusively large usage of TLSv1.2 and HTTP. Conversely, the partially correct prediction of User ID 7, User ID 47, and User ID 52 are mainly because of their unique protocols (SSLv3, PKIX-CRL, and DRBD) that served as discriminative features for accurate mapping of the corresponding packets to them.

TABLE 8. Top Three Most Frequently Used Port Sources

Port Source	Description	Number of packets	Percentage (%)
HTTPS (443)	HTTP protocol over TLS/SSL	53,084,748	55.41
HTTP (80)	World Wide Web HTTP	48,563,32	5.07
Pando-pub (7680)	Pando Media Public Distribution	958,047	1.00

TABLE 9. Unique Port Source

Unique Port Source	Description	Corresponded user	Number of packets
authentx (5067)	Authentx Service	User 2	60,112
13581	The default port for the SPX remote management service	User 2	57,523
13793	Unassigned	User 2	51,533

Based on Table 8, HTTPS and HTTP are the top two most frequently used Port Sources. HTTP uses port 80 to transmit unencrypted data between the users' browsers and WWW servers, whereas HTTPS uses port 443 to deliver encrypted data over a secured network. HTTPS port is more prevalent than HTTP port because the elevated concern toward Internet privacy has triggered a ubiquitous adoption of HTTPS as the secure version of HTTP (Naylor et al., 2014). The third most frequently used Port Source is Pando-pub (port 7680). This port is commonly used by WUDO (Windows Update Delivery Optimization) to distribute Windows updates using the peer-to-to-peer network connection. Its high prevalence is maybe because some users enabled the WUDO option to allow delivery of Windows updates from their devices to others on the Internet.

Table 9 shows some of the unique Port Sources (only the Port Source with the top three packets). The authentic (port 5067) exclusively used by User 2 is a registered port by Authentx Service that provides identity management, authentication, and credential issues solution. User ID 2 may have frequently used services or applications requiring credential authentication, like digital wallets and cryptocurrency. The other two unique ports of User ID 2 are ports 13581 and 13793. These ports are not officially assigned to any corporation or services by IANA (Internet Assigned Numbers Authority). However, Shadow Project SPX, a data recovery software, has set 13581 as the default port for its remote management service. Henceforth, we suspect that User ID 2 probably has installed this software on his device and routinely used port 13581 to monitor the progress of the data backup tasks.

TABLE 10. Top Three Most Frequently Used Port Destination

Port Destination	Description	Number of packets	Percentage (%)
HTTPS (443)	HTTP protocol over TLS/SSL	32,196,738	33.61
HTTP (80)	World Wide Web HTTP	2,669,004	2.79
56374	Dynamic Port	1,299,410	1.36

TABLE 11. Unique Port Destination

Unique Port Destination	Description	Corresponded user	Number of packets
argis-ds (2582)	ARGIS DS	User 2	80784
Authentx (5067)	AuthentX Service	User 2	78797
13581	The default port for the SPX remote management service	User 2	74664

Like Port Source (Table 10), the first two most prevalent Port Destinations are HTTPS and HTTP. As Port Destination works synchronously with Port Source, it is not surprising that they have this similarity. The third most frequently used Port Destination, 56374, is a dynamic port used temporarily by the client for communication with the server. It is also known as an ephemeral port that serves as a short-lived communication endpoint of the client that lasts only for the communication session between the client and server. Based on Table 11, the unique Port Destination with the top three most packets also showed similar port numbers to Port Source except for argis-ds (port 2582). Although it is a registered port assigned to ARGIS DS by IANA, its related application is relatively unknown.

TABLE 12. Decision Tree Analysis Using Port Source as Independent Variable

Observed	Predicted			
	User 1-33	User 34	User 35-96	Accuracy (%)
User 1-33	0	30,435,700	0	0.00
User 34	0	3,897,463	0	100.00
User 35-96	0	61,462,216	0	0.00
Overall percentage (%)	0.00	100.00	0.00	4.07

TABLE 13. Decision Tree Analysis Using Port Destination as Independent Variable

Observed	Predicted			
	User 1-33	User 34	User 35-96	Accuracy (%)
User 1-33	0	30,435,700	0	0.00
User 34	0	3,897,463	0	100.00
User 35-96	0	61,462,216	0	0.00
Overall percentage (%)	0.00	100.00	0.00	4.07

Both decision tree models generated using the Port Source and Port Destination (Table 12 and 13) as independent variables have identical accuracy with Destination IP. They encountered a similar technical issue in the decision tree analysis of Destination IP, where the tree-splitting stopped halfway when their categories exceeded the threshold of the CHAID growing method. Furthermore, CHAID could not identify any uniqueness of users other than User ID 34. User ID 34 achieved 100 % accuracy for his identification and contributed to 4.07% of overall model accuracy. He probably had a high degree of personality in his web browsing activities, enabling the correct mapping of all his packets.

The CHAID accuracy of the three decision tree models constructed using Destination IP, Port Source, and Port Destination separately is slightly lower than protocol. It is mainly due to the technical limitation caused by their excessive number of categories. When the number of attributes has reached a certain maximum, the recursive splitting of the decision tree will be automatically terminated (Milanović & Stamenković, 2016).

TABLE 14. Decision Tree Analysis Using All Independent Variable

User ID	Predicted (captured packets)									Accuracy (%)
	7	34	37	41	42	47	52	3, 57, 72	Others	
7	71,960	259,199	0	7,925	197,925	0	0	Trace amount	< 1	13.40
34	0	3,897463	0	0	0	0	0			100
37	0	402,876	19,912	45,010	466,641	0	0			2.06
41	2	279,616	0	377718	388,421	0	0			36.12
42	19	949,724	0	47,700	1,512,080	2	0			60.25
47	0	1,003,254	0	14,629	238,695	127625	0			9.22
52	1	664,141	0	160311	422,574	0	66,948			5.09
Others	21	77,669,045	1260	509,272	5,852,222	122,051	37			< 0.01
∑ %	0.007	88.86	0.002	1.21	9.48	0.026	0.07	< 1.00	0.00	6.36

All reliable independent variables - Destination IP and Protocol, Port Source, and Port Destination were used in combination for the decision tree construction (Table 14). The generated model has obtained an accuracy of 6.36 %. The protocol was the most significant variable to produce the first split of the parent node (user), followed by the Destination IP and then the Port Source, generating a tree with three levels of depth and 52 terminal nodes.

Although the model had excluded the Port Destination due to the termination of the tree growth, the created model still can identify a total of 20 users from their packets with varying accuracy. Again, User ID 34 achieved the highest accuracy (100%) with all his packets correctly mapped to him. As mentioned previously, he probably has a unique web browsing habit that enables the algorithm to discriminate his packets from other users. The other users, like User ID 42, User ID 41, and User ID 7, have their packets correctly predicted with accuracies of 60.25%, 36.12%, and 13.40%. It indicates that their web browsing patterns have some degree of personality but still share some common interests with other users. User ID 47, User ID 52, and User ID 37 also have a few packets identified accurately with percentages correct of 9.22%, 5.09 %, and 2.06%, respectively. 13 more users have trace amounts of correctly predicted network packets (less than 1%). They possibly had little uniqueness in their network packets, but the amount is insufficient to separate them from others.

Overall, the decision tree model generated with a combination of all the independent variables performs better than using each independent variable separately. It is probably because the combination of all four variables can promote the disorder of the data and result in a larger entropy that favours the discrimination of a user from others.

TABLE 15. Split-sample Validation Test

Training										
	Predicted (captured packets)									
User	7	34	37	41	42	47	52	13	Others	Accuracy (%)
7	57432	207225	0	6307	158226	0	0	Trace amount	< 1	13.38
34	0	3117709	0	0	0	0	0			100
37	0	322,422	15218	3,884	373388	0	0			2.06
41	2	223,491	0	302055	310626	0	0			36.12
42	13	759989	0	38189	1210420	1	0			60.26
47	0	802,850	0	11723	191110	102012	0			9.21
52	0	531473	0	128278	33079	0	53598			5.10
Others	18	62134046	1077	407444	4682986	97624	26			<0.01
∑ %	0.075	88.86	0.0021	1.21	9.48	0.026	0.007	<0.01	0.00	6.36
Test										
7	14,528	51,974	0	1,618	39,699	0	0	Trace amount	< 1	13.47
34	0	779,754	0	0	0	0	0			100
37	0	80,454	3,900	9,126	93,253	0	0			2.09
41	0	56,125	0	75,663	77,795	0	0			36.10
42	6	189,735	0	9,511	301,660	1	0			60.22
47	0	200,404	0	2,906	47,585	25,613	0			9.26
52	1	132,668	1	32,033	84,494	0	13,550			5.08
Others	4	15534999	249	101828	1169237	24427	4			<1.00
∑ %	0.080	88.87	0.02	1.21	9.47	0.26	0.07	< 1.00	0.00	6.36

A split-sample validation test (Table 15) was performed on the decision tree model with data partitioned into 80% training and 20% test samples. This validation test was used to examine its consistency and model reliability. Based on Table 16, both the decision tree models generated using the training and test sample have achieved identical accuracy (6.36%). There are only slight differences within the range of 0.1% for the accuracy in identifying some users. The result shows that the performance of the decision tree model is very consistent and precise, although the overall accuracy is not very high.

As the decision tree model's accuracy is low, future research requires further improvements. First, nearly 1 billion packet headers are involved in this research, and such a tremendous amount of data was challenging for the proposed statistical analysis. Large-scale data is often associated with irrelevant or unessential attributes that may interfere with the significant features and diminish the performance of the decision tree (Priyanka & Kumar, 2020). There are several possible ways to resolve this issue. One of the feasible methods is feature selection which extracts only relevant and significant features strongly associated with the user for subsequent classification (Clarke et al., 2017; Vinupaul et al., 2017).

In addition, grouping packets with similar characteristics by a clustering algorithm to create a more relevant feature can also reduce the total number of attributes and generate more valuable data (Miculan et al., 2019). Moreover, future research can implement other modified decision tree algorithms like the Size Constrained Decision Tree (SCDT) to maximise the classification accuracy under the size constraint of data. In contrast to a conventional decision tree, SCDT can control the increment of the leaf nodes and minimise the complexity of the tree through its dynamic combination of similar attributes (Wu et al., 2016). However, it is designed solely for binary splitting, so further modification to the meta-data is required to suit the user identification that involves multi-way splitting.

When dealing with a large-scale sample, the user identification based on their behavioural pattern of web browsing is often less accurate because many users may share similar interests (Yang, 2010). Based on the demography, the users in this research mainly come from a similar age range (22-24) and have identical educational backgrounds. As a result, the users in this research probably have similar web browsing patterns due to their common interests and routine activities. Therefore, we recommend participants from a more diverse background because users with varying ages and educations have different interests in their Internet usage.

Research on Malaysian university students shows that they mostly use the Internet for seeking information (30%), entertainment (23.2%), and education (19.4%) (Kurdus et al., 2017). On the other hand, secondary school students mainly use the Internet for researching homework (15.8%), playing games (12.7%), and listening to music (12.6%) (Ogur et al., 2017). Conversely, research in Oman found that adults above 40 rarely use the Internet for entertainment and social but mainly for their work-related activities (Khan et al., 2017).

Although the participants were requested to maintain their usual browsing habits, some probably altered their browsing habits or reduced their browsing frequency due to the concern of privacy. In addition, some of the users managed to capture more than three million packets despite some only contributing less than ten thousand packets. Although these variations are part of the users' distinctive browsing habits, they can also adversely affect the accuracy of the result because users with a relatively low frequency of packets are difficult to identify. Even though they possibly have some personalised web browsing activities, the amount of data is insufficient to highlight their uniqueness (Yang & Padmanabhan, 2010). Thus, a lengthened data collection period is vital to normalise the users' browsing habits. If future research can extend the data collection period to a week or month, the users are less likely to alter or reduce their web browsing habits. However, it may need to compromise with the increased data and participants' intention to enrol in the research.

CONCLUSION

This paper presented a classification method to generate a behavioural-based fingerprint capable of characterising users through their discriminative web navigation pattern. Among the four types of packet meta-data used to allocate the users' browsing behaviour, the protocol obtained the best outcome because it has relatively less insignificant and redundant attributes than the Destination IP, Port Source, and Port Destination. When using all these four features to match the users' packets with their identity, the acquired accuracy is slightly better than using each independently. This finding correlated with their larger information entropy generated for enhancement in the discriminating power of the digital fingerprint. Nevertheless, the accuracy of the created model still has vast room for improvement.

REFERENCES

Abdulazeez, A.M., Salim, B.W., Zeebaree, D.Q. & Doghramachi, D. 2020. Comparison of VPN Protocols at Network Layer Focusing on Wire Guard Protocol. International Journal of Interactive Mobile Technologies, 14(18): 157–177.

Adeniyi, E. 2019. Investigating The Factors That Promote Cybercrime Among University Students. Doctoral dissertation, Near East University.

Alotaibi, A.M., Fahaad Alrashidi, B., Naz, S. & Parveen, Z. 2017. Security issues in Protocols of TCP/IP Model at Layers Level. International Journal of Computer Networks and Communications Security, 5(5): 96–104.

Bakopoulou, E., Tillman, B. & Markopoulou, A. 2019. A Federated Learning Approach for Mobile Packet Classification. http://arxiv.org/abs/1907.13113 [10th, Jan 2022]

Buil-Gil, D., Miró-Llinares, F., Moneva, A., Kemp, S. & Díaz-Castaño, N. 2021. Cybercrime and shifts in opportunities during COVID-19: a preliminary analysis in the UK. European Societies, 23(S1): S47–S59.

Caviglione, L., Wendzel, S. & Mazurczyk, W. 2017. The Future of Digital Forensics: Challenges and the Road Ahead. IEEE Security and Privacy, 15(6): 12–17.

Cha, S. & Kim, H. 2017. Detecting encrypted traffic: A machine learning approach. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 10144 LNCS: 54–65.

Chen, J. & Cheng, W. 2016. Analysis of web traffic based on HTTP protocol. 2016 24th International Conference on Software, Telecommunications and Computer Networks, SoftCOM 2016.

Chen, P., Nikiforakis, N., Desmet, L. & Huygens, C. 2014. Security analysis of the Chinese web: How well is it protected? Proceedings of the ACM Conference on Computer and Communications Security 2014: 3–9.

Cheng, Y. C. & Wang, P. C. 2015. Packet Classification Using Dynamically Generated Decision Trees. IEEE Transactions on Computers, 64(2): 582–586.

Clarke, N., Li, F. & Furnell, S. 2017. A novel privacy preserving user identification approach for network traffic. Computers and Security, 70: 335–350.

Chan, D. 2021. Muhyiddin: Cyber security should be priority of every nation. https://www.nst.com.my/news/nation/2021/06/702934/muhyiddin-cyber-security-should-be-priority-every-nation [28th, May 2022]

Fadlil, A., Riadi, I. & Aji, S. 2017. Review of detection DDOS attack detection using naive bayes classifier for network forensics. Bulletin of Electrical Engineering and Informatics, 6(2): 140–148.

Ikuesan, A. R., Salleh, M., Venter, H. S., Razak, S. A. & Furnell, S. M. 2020. A heuristic for HTTP traffic identification in measuring user dissimilarity. Human-Intelligent Systems Integration, 2(1–4): 17–28.

Kathuria, M. & Gambhir, S. 2016. A Novel Optimisation Model for Efficient Packet Classification in WBAN. International Journal of Energy, Information and Communications, 7(4): 1–10.

Khan, M. A., Khan, S., Rehman, A. & Ghouse, S. M. 2017. Internet usage patterns: An exploratory study in Oman. International Journal of Applied Engineering Research, 12(7): 1232–1236.

Kurdus, N., Safiah, S., Zakiah, I., Massila, K., Abu Hassan, M. & Mohamed, S. 2017. Internet usage pattern and types of internet users among Malaysian university students. Journal of Engineering and Applied Sciences, 12(6): 1433–1439.

Laperdrix, P., Rudametkin, W. & Baudry, B. 2016. Beauty and the Beast: Diverting Modern Web Browsers to Build Unique Browser Fingerprints. Proceedings - 2016 IEEE Symposium on Security and Privacy. 878–894.

Malatras, A., Geneiatakis, D. & Vakalis, I. 2017. On the efficiency of user identification: a system-based approach. International Journal of Information Security, 16(6): 653–671.

Malaysian Communications and Multimedia Commission. 2020. Internet Users Survey 2020. The Internet Users Survey.

Meti, N., Narayan, D. G. & Baligar, V. P. 2017. Detection of distributed denial of service attacks using machine learning algorithms in software defined networks. 2017 International Conference on Advances in Computing, Communications, and Informatics, ICACCI 2017, 2017-Janua: 1366–1371.

Miculan, M., Foresti, G. L., & Piciarelli, C. 2019, February. Towards User Recognition by Shallow Web Traffic Inspection. ITASEC.

Milanović, M. & Stamenković, M. 2016. CHAID Decision Tree: Methodological Frame and Application. Economic Themes, 54(4): 563–586.

Mugisha, D., & Rughani, P. 2018. Web Browser Forensics: Evidence Collection and Analysis for Most Popular Web Browsers usage in Windows 10. Thesis in International Journal of Cyber Criminology.

Muzliza Mustafa. 2020. Two Malaysians Face Cybercrime, Money Laundering Charges in US — BenarNews. https://www.benarnews.org/english/news/malaysian/cyber-crime-09172020161753.html [17th September 2021]

Nath, K. 2015. Future What Comes after Web 3.0? Web 4.0 and the Future. International Conference on Computing and Communication Systems (I3CS'15), 1–4.

Ogur, B., Yilmaz, R. M. & Göktas, Y. 2017. An Examination of Secondary School Students' Habits of Using Internet. Pegem Journal of Education and Instruction, 7(3): 421–452.

Park, S., Jung, I. Y., Eom, H. & Yeom, H. Y. 2013. An analysis of replication enhancement for a high availability cluster. Journal of Information Processing Systems, 9(2): 205 –216.

Porter Felt, A., Barnes, R., King, A., Palmer, C., Bentzel, C. & Tabriz, P. 2017. Measuring HTTPS Adoption on the Web. Proceedings of the 26th USENIX Security Symposium, 1323–1338.

Priyanka & Kumar, D. 2020. Decision tree classifier: A detailed survey. International Journal of Information and Decision Sciences, 12(3): 246–269.

Rahim, R., Aryza, S., Wibowo, P., Harahap, A. K. Z., Suleman, A. R., Sihombing, E. E., Harputra, Y., et al. 2018. Prototype file transfer protocol application for LAN and Wi-Fi communication. International Journal of Engineering and Technology (UAE), 7(2.13 Special Issue 13): 345–347.

Raj S., V. & Gohain, M. 2021. Impact of Covid-19 on Malaysian E-Commerce. International Journal on Recent Trends in Business and Tourism 5(4): 8–10.

Ray, S. 2019. A Quick Review of Machine Learning Algorithms. Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Prespectives and Prospects, COMITCon 2019, 35–39.

Sandvine. 2019. The Global Internet Phenomena Report. Waterloo, Canada: Sandvine.

Santise, S., Cass, A. G. & Hall, S. 2012. Creating a Digital Fingerprint from Web Browsing History Alone.

Santise, S., & Cass, A. G. Creating a Digital Fingerprint From Web Browsing History Alone. http://orzo. u-nion. edu/Archives/Senior Projects/2012/CS. 2012/CSSenior Pro-ject Page-2012_files/Santise_Stephen_Report. pdf. [16th June 2021]

Sikos, L. F. 2020. Packet analysis for network forensics: A comprehensive survey. Forensic Science International: Digital Investigation 32: 200892.

Stuart Jacobs. 2016. Transport and Application Security Design and Use. Engineering Information Security, New Jersey, United States: John Wiley & Sons, Inc.

Vabalas, A., Gowen, E., Poliakoff, E. & Casson, A. J. 2019. Machine learning algorithm validation with a limited sample size. PLoS ONE, 14(11): 1–20.

Varadhan, S. 2016. Securing Traffic Tunnelled over TCP or UDP. Texas, US: Oracle Corporation.

Vinupaul, M. V., Bhattacharjee, R., Rajesh, R. & Kumar, G. S. 2017. User characterisation through network flow analysis. Proceedings of the 2016 International Conference on Data Science and Engineering, ICDSE 2016. India: Indian Institute of Technology, 1-6.

Wu, C. C., Chen, Y. L., Liu, Y. H. & Yang, X. Y. 2016. Decision tree induction with a constrained number of leaf nodes. Applied Intelligence, 45(3): 673–685.

Yang, Y. 2010. Web user behavioural profiling for user identification. Decision Support Systems, 49(3): 261–271.

Yang, Y. & Padmanabhan, B. 2010. Toward user patterns for online security: Observation time and online user identification. Decision Support Systems, 48(4): 548–558.

Zolfaghari, B., Srivastava, G., Roy, S., Nemati, H.R., Afghah, F., Koshiba, T. & Rai, B.K. 2020. Content delivery networks: State of the art, trends, and future roadmap. ACM Computing Surveys (CSUR), 53(2): 1–34.

Khairul Osman
T'ng Qi Feng
Hairee Izzam Mohd Noor
Noor Hazfalinda Hamzah
Gina Francesca Gabriel
Forensic Science Program, Faculty of Health Sciences (caw Bangi),
Universiti Kebangsaan Malaysia
khairos@ukm.edu.my, tngfenglz@gmail.com, a166341@siswa.ukm.edu.my,
drnoorhazfalindacsi@ukm.edu.my, ginafgabriel@ukm.edu.my