Human-Intelligent Systems Integration
https://doi.org/10.1007/s42454-020-00010-2

ORIGINAL PAPER

A heuristics for HTTP traffic identification in measuring user dissimilarity

Adeyemi R. Ikuesan[1] · Mazleena Salleh[2] · Hein S. Venter[3] · Shukor Abd Razak[2] · Steven M. Furnell[4]

Received: 14 May 2019 / Accepted: 18 May 2020

Abstract

The prevalence of HTTP web traffic on the Internet has long transcended the layer 7 classification, to layers such as layer 5 of the OSI model stack. This coupled with the integration-diversity of other layers and application layer protocols has made identification of user-initiated HTTP web traffic complex, thus increasing user anonymity on the Internet. This study reveals that, with the current complex nature of Internet and HTTP traffic, browser complexity, dynamic web programming structure, the surge in network delay, and unstable user behavior in network interaction, user-initiated requests can be accurately determined. The study utilizes HTTP request method of GET filtering, to develop a heuristic algorithm to identify user-initiated requests. The algorithm was experimentally tested on a group of users, to ascertain the certainty of identifying user-initiated requests. The result demonstrates that user-initiated HTTP requests can be reliably identified with a recall rate at 0.94 and *F*-measure at 0.969. Additionally, this study extends the paradigm of user identification based on the intrinsic characteristics of users, exhibited in network traffic. The application of these research findings finds relevance in user identification for insider investigation, e-commerce, and e-learning system as well as in network planning and management. Further, the findings from the study are relevant in web usage mining, where user-initiated action comprises the fundamental unit of measurement.

Keywords Heuristic algorithm · User-initiated HTTP request · GET method of HTTP request · Intrinsic network features · User inter-request time · User identification

1 Introduction

User identification transcends the branded tags of network and application layer identifiers of the TCP/IP stack. Challenges that surrounds unique identifiers, such as IP addresses, user security tokens, user IDs, and OS and browser fingerprinting, range from identifying the actual user behind the computer (especially in a multi-user single-computer system) to incorrect identification of user behind the computer. The reliance on unique identifiers at best gives a profile of the system and generic usage profile of the user, which further complicates the complexity of unveiling anonymity. The inherent nature of the Internet (in privacy and free-for-all paradigm) is the basis for which anonymity thrives.

Research on user identification in areas of behavioral inference (Fan et al. 2014; Yang 2010; Adeyemi et al. 2014), biometric dynamics (Ernsberger et al. 2017; Ikuesan and Venter 2018; Ikuesan et al. 2019), and network traffic analysis (Li et al. 2013a; Melnikov and Schönwälder 2010a; Adeyemi et al. 2016) are methods adapted for user identification through pattern extraction. The process employed for network

✉ Adeyemi R. Ikuesan
richard.ikuesan@ccq.edu.qa

Mazleena Salleh
mazleena@utm.my

Hein S. Venter
hein.venter@up.ac.za

Shukor Abd Razak
shukorar@utm.my

Steven M. Furnell
steven.furnell@nottingham.ac.uk

[1] Department of Cyber Security, Science and Technology Division, Community College of Qatar, Doha, Qatar

[2] School of Computing, Universiti Teknologi Malaysia, Skudai, Johor, Malaysia

[3] Digital Forensic Research Group, Computer Science Department, University of Pretoria, Pretoria, South Africa

[4] School of Computer Science, University of Nottingham, Nottingham, UK

Published online: 02 June 2020

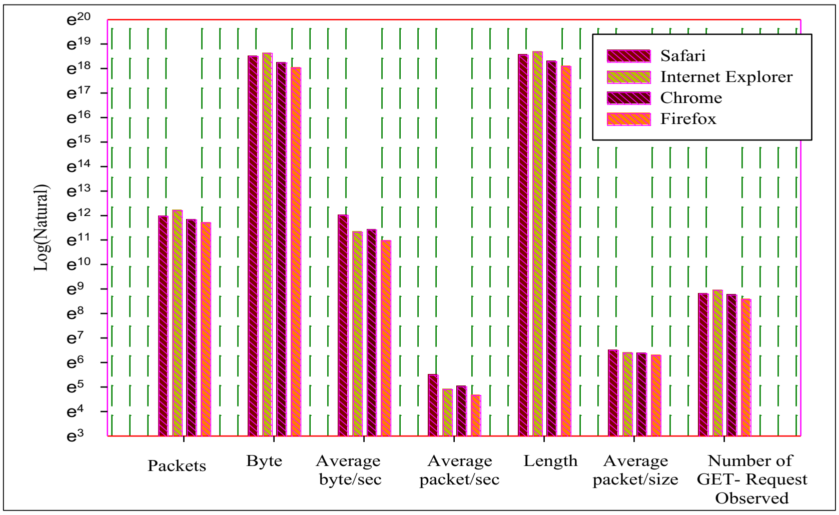
traffic pattern extraction includes logs and media scavenging, mining of audit trails, client-side caching, and extraction of flow records from captured network traffic. Additionally, cookie-based and keystroke-based authentication process (Mondal and Bours 2017; Mohlala et al. 2018; Iváncsy and Juhász 2007) are deployed for user identification. Such a method is similar to the integration of knowledge-based authentication into a behavior-based authentication process. A knowledge-based behavioral identifier system combines the uniqueness of user identifiers, user's unique pattern of usage, and other probable predefined characteristic behavior of users, to ascertain, to a degree of nearer certainty, the identity of a user. User profiling methods (Yang 2010; Adeyemi et al. 2016) attempt to establish facts based on the assumed/predefined salient features of the subject/object under observation. Similarly, such behavior-knowledge-based identifiers are applicable to network traffic profiling (Li et al. 2013a; Hlavacs et al. 1999).

However, the complexity of detecting a user-initiated HTTP web request (among several inline requests in addition to automated HTTP requests from system updates and network management processes) inhibits the effective implementation of online user profiling (Schneider et al. 2012). Furthermore, user-initiated requests differ statistically based on the type of user agent (such as a web browser) being used. Figure 1 illustrates a statistical dissimilarity observed among four commonly used web browsers. The statistical analysis in Fig. 1 is based on the exploration of twenty-one URLs for all browsers. The number of bytes, packets, and traffic frame length varies significantly from one browser to another, with Internet Explorer v10 generating the highest numbers. Safari generates the highest in terms of average bytes per second, average packet per second, and average packet size. This observation reveals the variation in the browser, similar to the empirical observation in (Schneider et al. 2012). Attempts to capitalize on TCP sessions or HTTP sessions as defined in (Hlavacs et al. 1999; Casilari 2001; Augustin and Mellouk 2011) further introduced complexities in the identification process of user-initiated requests. Such complexities include how to correctly identify user-initiated requests from a series of persistent/pipeline requests and how to establish accurate content-type header from HTTP body, as identified in (Schneider et al. 2012).

This study, therefore, attempts to develop a heuristic methodology that can cut through the estimation biases in HTTP traffic modeling, through empirical observation and statistical correlation of actual user-initiated requests. To this end, the study investigates the probability of user identification by distinguishing deliberate and or erroneous user-initiated requests. To the best of our understanding, this is the first study that presents user-initiated HTTP traffic extraction, which can be directly applied to practical web usage profiling, and user behavior modeling. This approach falls within the general knowledge of intelligent technology and analytics, as well as the complexities of the human and artificial systems, all within the broad domain of the intelligent systems integration. The implementation of the proposed approach presents a medium for attributing the event to a real human, thus demystifying the complexities associated with a typical online-related attribution problem.

The remainder of the paper is structured as follows: Section 2 gives a concise review of existing literature on user identification, with a special focus on network traffic usage. Various other heuristic methodologies for network traffic analysis are also discussed. The methodology through which the heuristic process for this study is designed is discussed in Section 3. Section 4 then discusses the detail of the algorithm, explicating its accuracy and reliability. Section 5 discusses the process of application of user inter-request time to distinguish different users.

Fig. 1 Statistical comparison of HTTP traffic under different web browsers



2 Related work

User identification in network traffic (Yang 2010; Melnikov and Schönwälder 2010a; Bhukya and Banothu 2011) has evolved from various perspectives. The study presented in (Bergadano et al. 2002) applied the keystroke dynamics of users for identification. This paradigm is also considered in (Bhukya and Banothu 2011) which extended the identification process to include behavior biometrics such as mouse movement pattern, keyboard dynamics, rotation of the wheel, and rate of keys pressed. Yang (Yang 2010) further extend this process by integrating user-centric features garnered from profiling web browsers. Empirical studies in (Schneider et al. 2012; Shaikh and Fiedler 2012; Velmurugan and Mohamed 2012) through a heuristic method revealed that embedded objects constitute the bulk of HTTP requests (involving web browsers). These heuristic methods either target performance analysis of network for network management (Schneider et al. 2012; Velmurugan and Mohamed 2012; Wang et al. 2011; Murta and Dutra 2004), dynamics of web usage for web management (Shaikh and Fiedler 2012; Nguyen 2005), comprehension of sources of network traffic (Hlavacs et al. 1999), or identification of web traffic (Augustin and Mellouk 2011; Chung et al. 2012; Charzinski 2000). The probability of identifying the user using the statistical correlation of network traffic of different users is proposed in (Melnikov and Schönwälder 2010a; Woods et al. 2012). The studies adopt HTTP flow duration to distinguish different users using plots of the cross-correlation function of flow duration to byte count and byte counts to packet counts of an HTTP flow. A study in (Lee and Gupta 2007) developed a heuristic process for HTTP traffic response arrival time analysis. The study interpolates timestamp, elapsed time of HTTP session, content length, and maximum time unit (MTU) of 62,000-client web proxy user request data, collected for 3 days. The study, however, assumed that the HTML object observed in a session, irrespective of the form of generation (user-initiated or system automation), is the start of the boundary of the user-initiated request. Using the threshold value of 1 s, from a support range of 30 s, a study in (Mah 1997) developed another form of heuristic process for HTTP connection as follows:

- If S_c = arrival time of start packet of connection-C_i
- E_c = arrival time of end packet of connection-C_i

C_1 and C_2 belong to the same connection if and only if the difference between a successive start packet arrival time (S_{c2}) and the end packet arrival time of previous connection (E_{c1}) is less than the defined threshold of 1 s (Sc_2-$Ec_1 \leq T_{\text{threshold}}$). Conversely, if $S_{c1} < S_{c2} < E_{c1}$, the connection is not the same. The study used Net Navigator as the user agent. However, the study did not consider pipelining and persistent connections. In addition, Net Navigator is an outdated web browser. A more recent study (Shaikh and Fiedler 2012) adopts the ON/OFF model similar to (Mah 1997), while studies in (Augustin and Mellouk 2011; Wang et al. 2011; Murta and Dutra 2004; Callahan et al. 2010) proposed heuristics centered on network traffic features and the GET-HTTP request method. Elucidation of the ON/OFF time model for traffic analysis is detailed in (Lee and Gupta 2007). Cuing from an economic model, a study in (Nguyen 2005) proposed a user action extraction model from the theory of demand and supply of user downloads. The study defined user action by a click on the hyperlink, or through an inserted URL. However, a study in (Iváncsy and Juhász 2007) opines that behavior profiling using only extrinsic traffic characteristics limits the potential of identifying user beyond the computer, since multiple users can use a single system within a comparable time window, and the converse is also true. The study, therefore, introduced a cookie-based authentication process in addition to web usage, where each user is assigned a unique identifier, which the study termed first-party and third-party cookie. The integration of such network traffic characteristics into an online identification process has recently gained wider adoption in online identification process (Li et al. 2013b). Studies in (Yang 2010; Yang and Padmanabhan 2010; Padmanabhan and Yang 2007) explored the process of online identification using network traffic behavior. The findings in (Padmanabhan and Yang 2007) assert that online browsing exhibit a fingerprint pattern that can be distinguished among online users. This is further examined by recent findings in (Pretorius et al. 2018; Adeyemi et al. 2017), where findings on the probability of online behavioral patterns of the individual user are further asserted. Online attribution using network traffic behavior thus presents a complementary platform for user authentication.

From the analysis so far, the dynamic characteristics of the user are not covered in any of the studies. This is relevant since a significant amount of network traffic generation is a function of the action of the user (Nguyen 2005). Furthermore, while these studies provide substantial empirically verifiable user identification techniques, the underlying question of "are we sure this traffic is user-initiated" remains unanswered. This study, therefore, focuses on the practicality of identifying an actual user-initiated request from network traffic, filling the research gap in existing heuristic methodologies.

3 Theoretical background

Unsurprisingly, given the popularity of the web, HTTP traffic type constitutes the highest volume of traffic observable in a TCP/IP network traffic (Schneider et al. 2012; Bergadano et al. 2002). Empirically, HTTP traffic exhibits patterns that are statistically significant in correlation with user behavior based on its keep-alive techniques (HTTP 1.0) and persistent connection (HTTP 1.1) that utilizes time-out process (Casilari 2001) which is also characterized by flexibility and interoperable property (Augustin and Mellouk 2011). A study in (Casilari 2001) asserts that there is no statistical difference between HTTP/1.0 and HTTP/1.1. A study in (Lee and Gupta 2007) asserts that the inter-request time of 1 s is not realistic, as well as the use of options field in HTTP header, as against assertion in (Nguyen 2005) which considers 1 s for an inter-request time. HTTP traffic is a typical client-server communication process that uses various safe methods such as GET, HEAD, POST, DELETE, PUT, TRACE, CONNECT, PATCH, and OPTION request response process (Fielding et al. 1999).

GET and HEAD request methods are conventionally defined for the retrieval process only. Typical HTTP request exhibits an empirically verifiable variation, which can be modeled using ON/OFF traffic model (Hlavacs et al. 1999; Shaikh and Fiedler 2012; Lee and Gupta 2007; Chen et al. 2008). ON/OFF model is a temporal model, which can be used to extract inherent characteristics from an object or subject under observation. The GET method of a web request is one of the prescribed request methods for HTTP client-server communication (Fielding et al. 1999; Choi and Limb 1999). The current study builds on the heuristics developed in (Nguyen 2005; Lee and Gupta 2007; Mah 1997; Choi and Limb 1999), in which a web request is used as the basic unit for user behavior inference, through the metadata in HTTP header.

Choi and Limb (Choi and Limb 1999) indicate that user-generated web request is not a function of the web-caching (a process that uses the if-modified-since option field of the HTTP header). Furthermore, their study describes a methodology for user-generated HTTP request as:

- The first object in a request is an HTML document, that is, a request for an object whose extension is either ".htm, .asp, .cgi" or a URL that ends with '/', and or a request whose response is a MIME type "text/HTML", and the status code is 200 (OK).

The heuristic did not specify the number of "/" a request URL could contain. Furthermore, the limitation of object extension to only three formats makes the heuristic less generic. Additionally, study in (Choi and Limb 1999) describes hyperlink click that results in a single object download such as image (.jpg or .pnp), text (.pdf or .doc), and sound (.avi or .mp3) files as inline object. This exclusion further reduces the potential of the methodology from accurately identifying a user-initiated web request. From empirical observation, it is erroneous to classify all multimedia attributes of HTTP request that generate a single object request as an inline request, when the user deliberately requested for such a page.

Besides, a single document file (namely text document in pdf, doc, ppt, and other related file formats) are not embedded into web pages but are stand-alone elements in a page. Hence, this study included the complexity of a single object request into the boundary of the user-initiated request. Such a method would also require the use of a heuristic methodology. Heuristic-based intelligent processes provide simple logical metrics for understanding a complex phenomenon (Illankoon et al. 2019). Such simplicity is essential in understanding patterns generated from human action.

Figure 2 shows the theoretical description of the HTTP web request structure. The request for a single base file generates tens of inline files. A closer examination of the base file and the inline files reveals a seemingly distinct difference between the two types of web files. Base files (such as "https://www.google.com/, https://facebook.com/, https://wikipedia.com/") typically have a singular "/" appended at the end of the identifier of the URL. It is, however, dissimilar to the number of "/" for an inline object, which could range to tens, relative to the base file. It should, however, be mentioned that some base files (e.g., "https://en.wikipedia.org/wiki/Main_Page") may have ≤ 2 number(s) of "/" depending on the dynamics of the URL identifiers. Hence, a logical borderline of the base file and inline file is in the deterministic approximation of a number of "/" in a given user-initiated URL. Moreover, this generalization does not distinguish bots and other forms of zombies on the network. This study, therefore, defines user-initiated web request to encompass a more generic description:

- an HTTP request (method = GET). Other forms of HTTP request methods such as POST and HEAD are not considered in this study because such methods require higher security clearance level to initiate, which is not usually available to public users.
- a URL with ≤ 2 number(s) of "/"and character comparison where two consecutive identified request of less than or equal to 1/3 similarity index.
- Consecutive URL with similarity greater than 2/3 must have time difference (Δt) of more than 8 s.

As illustrated in Fig. 3, this process passes through a filter rule as a preprocessing stage, followed by heuristic pattern observation. The heuristic process considers file format, the unique signature generated by a user-initiated request, and the time difference condition.

4 Intelligent-heuristic methodology

As evident in Fig. 1, the ratio of the actual GET request to the observed Get request differs from one web browser to another

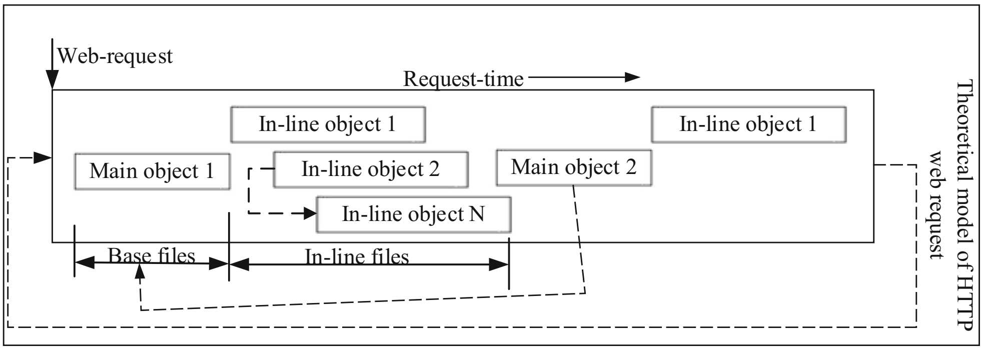


Fig. 2 Theoretical composition of HTTP web request

(approximately 3e^{-3}, 2.6e^{-3}, 3.2e^{-3}, and 3.8e^{-3} for Safari, Internet Explorer, Firefox, and Chrome browsers respectively). This observation is also supported by the assertion in (Schneider et al. 2012; Lee and Gupta 2007; Casilari et al. 2001) which states that the bulk of HTTP web requests is not human-initiated. The current study is based on the principle of user dynamics as reflected in the intrinsic characteristic pattern found in a typical client-server communication system. Unlike web page pattern identification, the user request pattern is a complex structure, which exhibits patterns based on user action. The logic behind heuristic methodology for user initial request pattern identification is similar to the studies in (Shaikh and Fiedler 2012; Lee and Gupta 2007; Choi and Limb 1999). The filter rule and observed signature of user-initiated requests distinguish the heuristic developed in this study from existing literature. Features observed include source port, frame length, requested URL, the frame number of the request as detailed in Table 1. These features were considered based on the initial observation of the experimental process, where it was observed that these features represent the principal composition of human-initiated actions.

The observed features are logically meaningful parameters (Mah 1997) which capture probable user requests from all layers of the TCP/IP stack. Filter rule (ip.src==xx.xx.xx.xx) and ((http.request.method==GET)) and a comma-separated value (csv) file output forms the preprocessing stage. Figure 4 shows the pseudocode for the heuristic process. The filtered file is structured in a table of columns and rows, where the columns are the observed features and each row depicts a request entry.

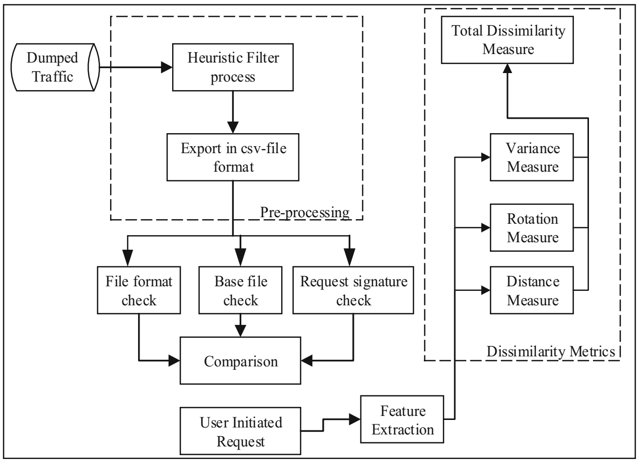


Fig. 3 Process of user-initiated request identification and dissimilarity metrics

The process begins by polling for HTTP requests which have the same source port number of ≥ 4 consecutive values. The ≥ 4 consecutive source port values were defined based on the empirical observation of the distribution of human-initiated requests. The first row of the identified sequence is tagged ith row, while subsequent rows are tagged in increasing order of 1 (⇒ $i_{\text{next\_value}} = i\text{th} + 1$). The next sequence involves verifying if the identified ith row is a user-initiated request. The frame-length column is tagged as the jth value. The ith and $i_{\text{next\_values}}$ row of "j-column" are compared with the sequential value of ± 1. This sequence is then followed by another round of verification. In line 8 to line 16, the process checks if the first and any other observed URL encountered in the sequence is a user-initiated URL based on the acceptable file format using a look-up table. In line 17, the process compares 4 consecutive source port and the corresponding frame length. If the condition is met, then the probability of the request being human-initiated is weighted 1/4R. From line 17 to line 35, the algorithm adapts the logical combination of features from the lower level of the TCP/IP model stalk.

At line 36, the process moves to a higher level on the TCP/IP model: the application layer. From line 37 to line 40, the process examines the URL of the appended web request for the number of "/". This is considered based on the defined threshold of ≤ 2 numbers of "/". The fulfillment of these criteria results in an additional weighted probability of 1/4R value. The aggregation of these weighted probabilities results in 1/2R overall weighted probability. This implies that the appended URL has a 50% likelihood of being user-initiated. From line 42 to line 47, the process considers two consecutive URL using the initially appended URL as the reference point. It compares the appended URL with the preceding URL based on the defined similarity index. The satisfaction of these criteria improves the weighted probability by 1/4R weight. The aggregation at this stage gives a 75% probability of being

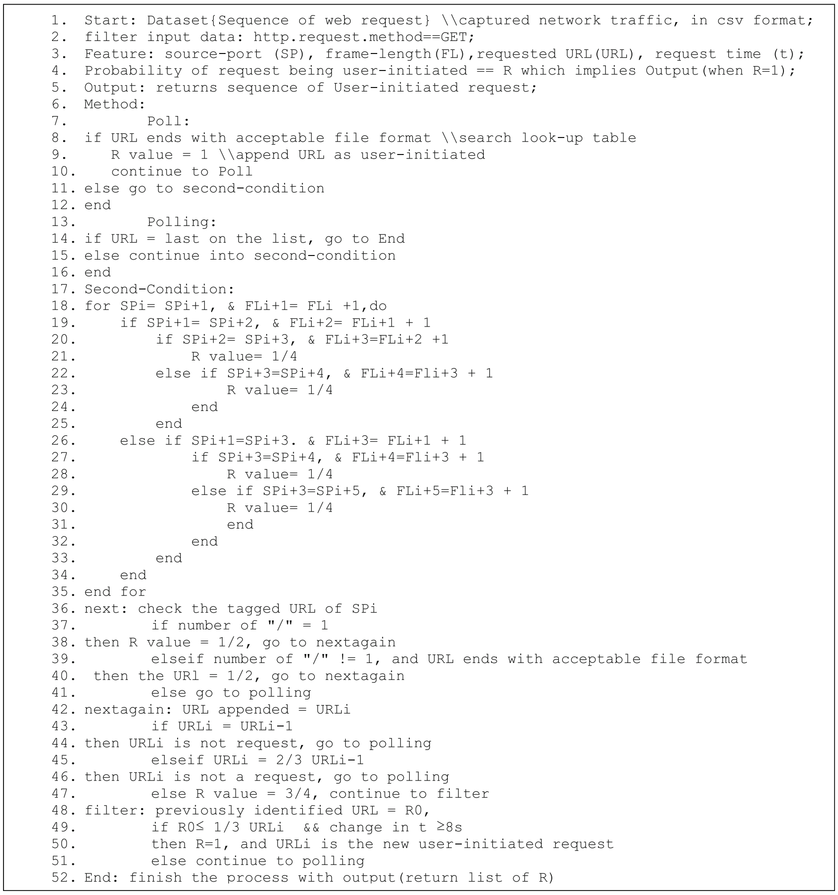
Table 1 Features and description

Features	Description
Source port	The port number of the source of the request (in line with the source IP address)
Get method of HTTP request	GET method of typical HTTP web request
Request frame length	HTTP request length similar to study in (Mah 1997) is the observed length of the HTTP request filtered through the HTTP GET request which is the actual length of the web request captured on the network interface card
Visited URL	The observed uniform resource locator address requested by the user
Request time	The actual time requested resources as captured at the network interface card
Request frame number	Sequential ordering by time of observed HTTP request frame as captured at network interface card
TCP SYNC flag	TCP SYNC flag bit is set. One of three stages of the TCP 3-way handshake process of TCP connection establishment

a user-initiated request. The final verification process spans line 48 to line 50. It considers the actual request time of the appended URL for a time lag threshold. It examines the appended URL with the preceding URL for a time lag of ≥8 s. In the experiment, 8 s was observed to be the minimum delineating time lag from one user-initiated request to another (hereinafter referred to as think time). This variable (think time threshold of 8 s) can be dynamically defined based on the behavior of the user, and network delays. The fulfillment of this time-based criterion amounts to the unity of probability weight ($R = 1$). At this stage, the observed request is defined as user-initiated with a probability of 1.

4.1 Ground truth for data analysis

An experimental process was conducted with a group of users. The experimental process was conducted in a laboratory where all workstations have a network and screen capture tools installed (Wireshark, and CamStudio in this case). The participants were briefed on the research ethics and the



```
1.  Start: Dataset{Sequence of web request} \\captured network traffic, in csv format;
2.  filter input data: http.request.method==GET;
3.  Feature: source-port (SP), frame-length(FL),requested URL(URL), request time (t);
4.  Probability of request being user-initiated == R which implies Output(when R=1);
5.  Output: returns sequence of User-initiated request;
6.  Method:
7.         Poll:
8.  if URL ends with acceptable file format \\search look-up table
9.     R value = 1 \\append URL as user-initiated
10.    continue to Poll
11. else go to second-condition
12. end
13.        Polling:
14. if URL = last on the list, go to End
15. else continue into second-condition
16. end
17. Second-Condition:
18. for SPi= SPi+1, & FLi+1= FLi +1,do
19.     if SPi+1= SPi+2, & FLi+2= FLi+1 + 1
20.         if SPi+2= SPi+3, & FLi+3=FLi+2 +1
21.             R value= 1/4
22.         else if SPi+3=SPi+4, & FLi+4=Fli+3 + 1
23.                 R value= 1/4
24.             end
25.         end
26.     else if SPi+1=SPi+3. & FLi+3= FLi+1 + 1
27.             if SPi+3=SPi+4, & FLi+4=Fli+3 + 1
28.                 R value= 1/4
29.             else if SPi+3=SPi+5, & FLi+5=Fli+3 + 1
30.                 R value= 1/4
31.                 end
32.             end
33.         end
34.     end
35. end for
36. next: check the tagged URL of SPi
37.        if number of "/" = 1
38. then R value = 1/2, go to nextagain
39.        elseif number of "/" != 1, and URL ends with acceptable file format
40.  then the URl = 1/2, go to nextagain
41.        else go to polling
42. nextagain: URL appended = URLi
43.        if URLi = URLi-1
44. then URLi is not request, go to polling
45.        elseif URLi = 2/3 URLi-1
46. then URLi is not a request, go to polling
47.        else R value = 3/4, continue to filter
48. filter: previously identified URL = R0,
49.        if R0≤ 1/3 URLi  && change in t ≥8s
50.        then R=1, and URLi is the new user-initiated request
51.        else continue to polling
52. End: finish the process with output(return list of R)
```

Fig. 4 User-initiated request identification pseudocode

installed software on the computer. The consent form was distributed to all participants, prior to the experimental process. A total of 17 respondents participated in this study. The participants were given a printed list of URLs that could be visited, an e-copy of some other possible URL, and other possible information searchable on the Internet. To develop a scientific ground truth of the expected dataset, each participant was asked to give a written record of the number of URLs visited, and the number of search information (file formats) searched. The installed screen capture tool is used to ascertain the claims being made by the participants as shown in Table 2. Six (6) participants (hereinafter defined as a user) provided a complete record of their activities, while other users (11 participants) had truncated screen capture, truncated network traffic capture, or no written record of activities.

Though the actual number of users adopted for this study is relatively small compared with other online identification studies (Padmanabhan and Yang 2007; Adeyemi et al. 2017), it, however, satisfies the purpose of the study, which is to validate the reliability of the heuristic algorithm, as well as to analyze network traffic for observable patterns in a user-initiated request. Thus, the study assumes that the measure of the statistical significance of the sample size does not apply to this study. This is because the study aims to measure the reliability and accuracy of the proposed algorithm, which is independent of the sample size. However, the sample size is also similar to the sample size used in a cybernetic study presented in (Melnikov and Schönwälder 2010a), where only six respondents were investigated. Furthermore, the size of the dataset provides a sufficient URL request sample suitable to validate the developed heuristic algorithm and to explore user inter-request time. We observed a wide variation between the record of the participants and the actual screen captured as shown in Table 2.

In order to clarify this disparity, and ensure the accuracy of the selected sampled respondents, the study manually verified the request of each user using the data from the captured screen. The efficiency of the algorithm is verified using the *F*-measure of a weighted average of recall and precision. *F*-measure ranges from a very poor score of 0, to a very efficient score of 1. The mathematical representations of recall rate, precision level, and *F*-measure are given in Table 3.

4.2 Methodology for user traffic pattern dissimilarity measurement

The user-initiated request extraction process defined in the previous section is structured to identify the start of the precise request of the user. For each user, the two-unit vector (request and inter-request time) is formed. The *i*th request time of a user, and the row, *j*, represents the time between time t_{j-1} and t_j of the user request time (also referred to as user think time). Think time reflects the disposition of the user at that point in time, which when aggregated over a given duration, reflects the personality of the user (Zhang 2006). Request and inter-request time of the user-initiated request form the matrix for each user. The choice of these features is to reflect the distinct nature of users, based on the assertion that an individual request pattern is dependent on the user, and not on the systems under investigation.

To determine the dissimilarity between two users, this study adopts the metrics identified in (Parthasarathy 2005), which developed a flexible formulation of metric variation through a linear combination of aggregated metrics. Distance, rotational, and variance measure constitute the dissimilarity metrics. This study implemented these measures as articulated mathematically in (Lakhina et al. 2004) using Matlab software. The essence of the adopted measures is as follows. Distance measure (D_d) reflects the intrinsic characteristics of the dataset in terms of its differences to its centroid. The rotational measure (D_r) gives the structural differences in the datasets, by revealing its latent demography through structural transformation. The variance measure (D_v) reveals the shape dissimilarities of the dataset.

Euclidean distance measure (Ekström 2011), singular value decomposition (Vedral 2002), and Kullback-Leibler divergence also referred to as the theory of relative entropy (Klema and Laub 1980), which incorporates principal component analysis and variance, are used for the distance, rotational, and variance measures respectively. Overall dissimilarity is measured using the equations shown in (1) and (2).

$$\text{Generic dissimilarity (GD)} = D_\mathrm{d} \times D_\mathrm{r} \times D_\mathrm{v} \tag{1}$$

Table 2 Ground truth formulation for the dataset

Users	Process	User record	Screen capture record
1	• Please browse as many URLs as possible from the electronic list given • Please browse as many URLs as possible from the printed listed given • Please search for any document of interest relating to your study on the Internet • Please search for any two or more document files relating to user authentication, on the Internet.	21	17
2		14	16
3		42	38
4		22	19
5		25	108
6		25	49

Table 3 Evaluation metric for heuristic efficiency

Evaluation Method	Metrics	Description
Accuracy evaluation	*F*-measure	$F = 2 \times \frac{\text{precision.recall}}{\text{precision+recall}}$
	Recall	$\text{recall (sensitivity)} = \frac{\text{TP}}{\text{TP+FN}}$
	Precision	$\text{precision} = \frac{\text{TP}}{\text{TP+FP}}$

TP true positive, *FN* false negative, *FP* false positive

$$\text{Weighted measurement dissimilarity (WMD)}$$
$$= \beta_0 + (\beta_1 \times D_\text{d}) + (\beta_2 \times D_\text{r}) + (\beta_3 \times D_\text{v}) \qquad (2)$$

Equation (1) expresses a pairwise dissimilar pattern based on the composite effect of the structural, shape, and distance measure. Equation (2) on the other hand gives a weighted pairwise effect peculiar to the measure of interest, by equating the coefficient of the desired dissimilarity to 1, and others to 0.

5 Result and discussion

This section presents the result of the accuracy and the sensitivity of the developed heuristic algorithm and the dissimilarity measure for user request and inter-request time. The acquired data set (network traffic) is first passed through a filter to ensure that the traffic originated from the sampled users, as discussed in the previous section. To understand the statistical characterization of the temporal-spatial structure of the inter-request and request time of the users, the study examines a simple plot of the filtered output dataset as shown in Fig. 5. Visual observation of the plot reveals a dissimilar pattern between user5, user1, and other users. Structural dissimilarity among other users is relatively visually oblique.

Table 4 shows the result of the reliability of the developed algorithm on the experimental test set. The result shows that the algorithm exhibits efficient output (the outcome of the *F*-measure test). The output of the algorithm serves as the input to the dissimilarity measure as illustrated in Fig. 2. Tables 5, 6, and 7 present the findings on the various dissimilarity measure adopted for this study.

In the D_d as shown in Table 5, the overall dissimilarity index at the inter-request time of user feature map is higher than the request-time user feature map. The relationship between [user1 and user5] and [user2 and user3] is observed to have a relatively higher correlation index based on the request time. [User3 and user6] and [user2 and user5] are observed to exhibit a higher correlation at the inter-request-time map. However, the result is different from other interactions at both

Fig. 5 Plot of user request time

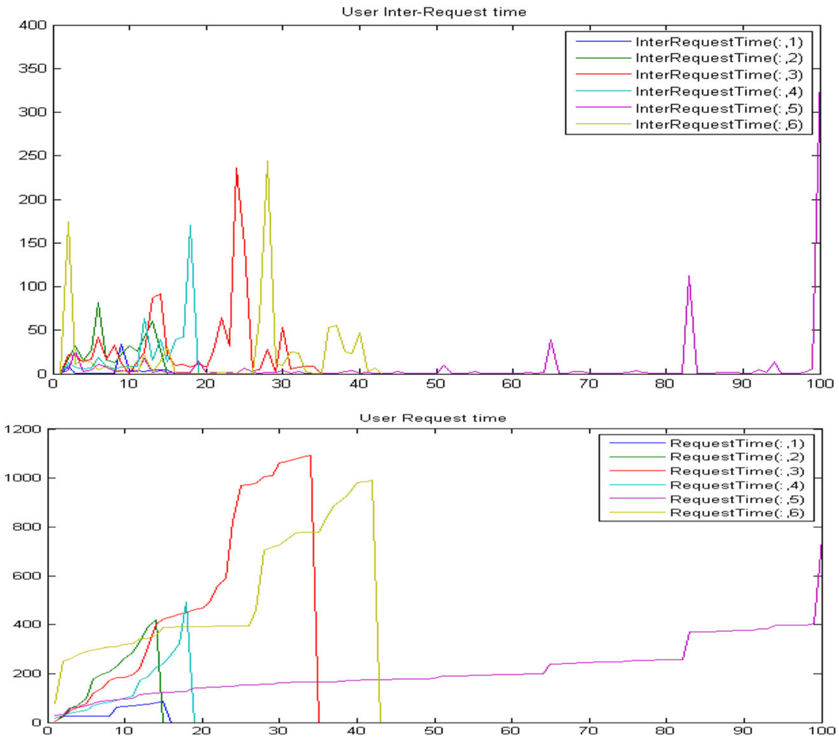


Table 4 Result of the efficiency of the heuristic algorithm

Users	FP	FN	TP	Recall	Precision	*F*-measure
1	0	1	16	0.9412	1	0.9697
2	0	1	15	0.9412	1	0.9697
3	0	1	37	0.9412	1	0.9697
4	0	1	18	0.9412	1	0.9697
5	0	1	107	0.9412	1	0.9697
6	0	3	46	0.9388	1	0.9684
Average				0.9408	1	0.9694

Table 6 Rotational measure

	User1	User2	User3	User4	User5	User6
User1	0	0.1113	0.2056	0.2883	0.1215	0.1894
User2	0.1113	0	0.0942	0.3997	0.2328	0.0780
User3	0.2056	0.0942	0	0.4939	0.3270	0.0162
User4	0.2883	0.3997	0.4939	0	0.1668	0.4777
User5	0.1215	0.2328	0.3270	0.1668	0	0.3108
User6	0.1894	0.0780	0.0162	0.4777	0.3108	0

the request time and the inter-request time. It also reveals the relatively stronger dissimilarity aggregate among the users.

The user request time is observed to have a lower dissimilarity weight relative to the inter-request time. This is expected since the experiment was conducted within the same hour, and each user could be said to operate at the almost closer time clock, which is considered as the time stamp for the user-initiated request.

Rotational and variance measure, on the other hand, gives the discrete components of the request and inter-request time. The D_r as shown in Table 6 reveals a mixture of lower similarity and more dissimilarity index between different users. User6 shows stronger similarity with user2 and user3, while others exhibit a varying degree of dissimilarity. This reveals further that the rotational measure, though produces lower index weight, can be a useful tool in dissimilarity measure. The D_v, also referred to as the shape parameter as shown in Table 7, reveals more similarity index than the rotational measure. User1 shows relative similarity with user2 and user5, while user2 shows stronger similarity index with user3 and user6.

The overall dissimilarity measure shown in Table 8 reveals the impact of the two features as well as the level of the affinity between the users. From the inter-request-time map, user1 relates dissimilarly with user3, user4, and user6. User4 shows dissimilarity with user1, user3, and user6. User5 shows dissimilarity with user3 and user6. From the request-time map, more similarity correlation is shown among the users. In general, user5 exhibits stronger dissimilarity with other users, as evident in the higher-level abstraction shown in Fig. 4. These results thus corroborate the result that dissimilarity measures can extract higher-level abstractions, as well as the statistical dissimilarity. Table 9 gives a concise description of the users that share a similar pattern (as highlighted).

The similarity strength observed in [user3 & user6] and [user2 & user3] cuts across the dissimilarity measure. This shows that while some users can be distinguished from others, they may also share a similar pattern with other users. Furthermore, the similarity strength between user2 and user6 is relatively higher, which also confirms the logic of the relationship between user3 and user6 with respect to user2. From the tabulated result, the following observation can be deduced.

i. Inter-request time is a useful feature for user identification research
ii. User request time of HTTP traffic can be used to profile network users
iii. Inter-request-time and request-time time series data satisfy the distance metric criteria as identified in (Lin et al. 2012), which include symmetry, triangular inequality, constancy, and positivity.
iv. The designed algorithm works efficiently for identification of user-initiated requests

Table 5 Distance measure of user feature map

	Request time						Inter-request time					
	User1	User2	User3	User4	User5	User6	User1	User2	User3	User4	User5	User6
User1	0	629.41	757.19	478.83	*7.81*	299.06	0	22,250.19	220,804.24	9598.27	27,558.76	221,183.80
User2	629.41	0	5.90	10.28	497.02	283.70	22,250.19	0	102,869.75	2620.85	283.70	103,128.87
User3	757.19	5.90	0	31.75	611.23	104.53	220,804.24	102,869.75	0	138,329.98	92,349.01	0.16
User4	478.83	10.28	31.75	0	364.36	21.06	9598.27	2620.85	138,329.98	0	4629.11	138,630.43
User5	*7.81*	497.02	611.23	364.36	0	210.23	27,558.76	283.70	92,349.01	4629.11	0	92,594.54
User6	299.06	283.70	104.53	21.06	210.23	0	221,183.80	103,128.87	0.16	138,630.43	92,594.54	0

Table 7 Variance (shape parameter) measure

	User1	User2	User3	User4	User5	User6
User1	0	0.0062	0.0209	0.0432	0.0075	0.0178
User2	0.0062	0	0.0044	0.0837	0.0276	0.0030
User3	0.0209	0.0044	0	0.1281	0.0547	1.3131
User4	0.0432	0.0837	0.1281	0	0.0136	0.1088
User5	0.0075	0.0276	0.0547	0.0136	0	0.0473
User6	0.0178	0.0030	1.3131	0.1088	0.0473	0

v. Some user exhibits a similar pattern of think time. Uncovering the underlying reason behind such occurrence could be a new direction for research into user identification

vi. Dissimilarity measures present a multi-view orthogonal projection of the structure and pattern in the dataset, which reflects the usage pattern of each user. These patterns show the structural and statistical characteristics inherent in the data.

vii. There exists a common consensus (convergence) on the dissimilarity of some users, in this case, [user2 and user3] and [user3 and user6].

Given these observations, this study thus supports the assertion in (Padmanabhan and Yang 2007; Melnikov and Schönwälder 2010b) that the online behavior of individual user exhibits pattern that differs relatively from another user. The first and second observations expound on the strength of web request characteristics in distinguishing among online users. The third observation highlights the robustness and efficiency of the proposed heuristic method. Also, the result presented in this study shows that user dissimilarity, which is based solely on human-initiated action, provides substantial dissimilarity weight for online user observation. Contrasting these findings to observation in network traffic management research such as in (Salleh et al. 2005; Guan et al. 2010), where network traffic distribution is observed to follow the Poissonian distribution, this study reveals that, in order to explore online identification using network traffic (particularly from the client side), it is essential to extract only human-initiated action from the bulk of network traffic data. This assertion is further supported by the fifth and sixth observations. It is, therefore, logical to assert that the application of such a heuristic process into online identification techniques presents a platform for an effective online identification process.

With respect to intelligent systems interaction, the developed heuristic can be used to develop a context-free autonomous web profiling system. Typical web profiling systems, such as recommender systems, rely on a series of behavioral attributes to establish reliable patterns upon which profiling is then built. However, such profiling systems are not suited for a multi-user single-system scenario. Furthermore, a complex interactive mechanism would be required to profile users in such a scenario. The application of the developed heuristic algorithm, combined with the second observation from this study, can provide a substratum upon which a reliable intelligent user profiling system can be developed. An intelligent system capable of providing a reliable method of attributing a given action on the web, to a known use, is a major requirement in typical recommender systems and e-commerce platforms. These platforms require a human-centered design which is satisfied by the developed algorithm in this study. Furthermore, the demystification of online identity complexities presents a medium for user attribution in security and digital investigations.

6 Conclusion and future work

The empirical validation presented in this study rejects the null hypothesis in favor of the alternate hypothesis, which states that user identification on the Internet is feasible, provided that deliberate user-initiated request can be identified and extracted. This study developed a novel heuristic algorithm that accurately identifies the user-initiated request of HTTP web traffic. The study further presents empirical findings on the

Table 8 Total dissimilarity metrics

	Request time						Inter-request time					
	User1	User2	User3	User4	User5	User6	User1	User2	User3	User4	User5	User6
User1	0	0.43	3.25	*5.97*	0.01	1.01	0	15.25	949.70	119.67	25.02	743.81
User2	0.43	0	0.00	0.34	3.20	0.01	15.25	0	42.89	87.66	*1.83*	24.41
User3	3.25	*0.00*	0	2.01	*10.93*	*0.00*	949.70	42.89	0	8753.37	1651.66	*0.00*
User4	*5.97*	0.34	2.01	0	0.82	1.09	119.67	87.66	8753.37	0	10.47	7201.34
User5	0.01	3.20	*10.93*	0.82	0	3.09	25.02	*1.83*	1651.66	10.47	0	1360.17
User6	1.01	0.01	*0.00*	1.09	3.09	0	743.81	24.41	*0.00*	7201.34	1360.17	0

Table 9 Summary of similarity against dissimilarity

Measure	Similarity strength	
D_d	Request time	User1 & user5, and **user2 & user3**
	Inter-request time	**User3 & user6** and user2 & user5
D_r	***User2 & user3*** and ***user3 & user6***	
D_v	User1 & user2, user1 & user5, **user2 & user3**, and user2 & user6	
GD	Request time	**User2 & user3**, user2 & user6, and **user3 & user6**
	Inter-request time	User2 & user5 and **user3 & user6**

relevance of user inter-request time and user request time in distinguishing between different users. To achieve this, the study adopts a multiplier effect of distance (Euclidean distance), rotational (angle between the principal component using singular value decomposition of the covariance matrix of the user data), and variance (shape parameter: Kullback-Leibler divergence of the principal component of the user data) measures as dissimilarity metrics. The result from the dissimilarity metrics reveals the intrinsic dimensionality of different users, from which the user identification process can be defined. Future research could consider extensive user dissimilarity process using the integration of methods such as autoregressive moving average, "Bag of pattern" representation, and dynamic time warping techniques. Such techniques can be applied to other protocols, from which generic assumptions about human online behavior can be inferred. Furthermore, attempts can be made to explore the application of user dissimilarity in a more generic perspective using the various psychosocial attributions. Such research will find relevance in unveiling user anonymity over the Internet for e-commerce, cyber forensic purposes, and network traffic modeling. As part of an ongoing research process, a broader scope of the website such as e-commerce web pages, e-learning web pages, and social networking web pages will be specifically considered to observe human-initiated action and the consequential behavioral pattern of individual users. Further findings from such findings will play a vital role in online profiling processes specific to social networking sites. The integration of these findings will be used to build an online profiling software for e-profiling purposes, as well as in law enforcement profiling processes. Particularly, the software will consider the integration of intelligence of the extended heuristics in its engine for profiling.

Funding Information Open Access funding provided by the Qatar National Library.

Compliance with ethical standards

The participants were briefed on the research ethics and the installed software on the computer. The consent form was distributed to all participants, prior to the experimental process.

Conflict of interest The authors declare that they have no conflict of interest.

References

Adeyemi IR, Razak AS, Salleh M (2014) A psychographic framework for online user identification. In: International Symposium on Biometrics and Security Technologies (ISBAST), pp. 198–203

Adeyemi IR, Razak SA, Salleh M, Venter HS (2016) Observing consistency in online communication patterns for user re-identification. PLoS One 11(12):1–27

Adeyemi IR, Razak SA, Salleh M, Venter HS (2017) Leveraging human thinking style for user attribution in digital forensic process. Int J Adv Sci Eng Inf Technol 7(1):198–206

Augustin B, Mellouk A (2011) On traffic patterns of HTTP applications. In: IEEE Global Telecommunications Conference(GLOBECOM), pp. 1–6

Bergadano F, Gunetti D, Picardi C (2002) User authentication through keystroke dynamics. ACM Trans Inf 5(4):367–397

Bhukya W, Banothu S (2011) Investigative behavior profiling with one class SVM for computer forensics. Multi-disciplinary Trends Artif Intell:373–383

Callahan T, Allman M, Paxson V (2010) A longitudinal view of HTTP traffic. Passiv Act Meas:222–231

Casilari E (2001) Modeling of HTTP traffic. IEEE Commun Lett 5(6): 272–274

Casilari E, González FJ, Sandoval F (2001) Characterisation of web traffic. In: In Global Telecommunications Conference, pp 1862–1866

Charzinski J (2000) HTTP/TCP connection and flow characteristics. Perform Eval 42(2–3):149–162

Chen C, Xu Y, Zhang L (2008) Some remarks on ON/OFF network traffic. In: Workshop on power electronics and intelligent transportation system

Choi H-KK, Limb JO (1999) A behavioral model of web traffic. In: Network Protocols, 1999. (ICNP'99) proceedings. Seventh International Conference, pp 327–334

Chung J, Li J, Choi Y, Hong J (2012) Application traffic identification based on remote subnet grouping. In: Network Operations and Management Symposium (APNOMS), 2012 14th Asia-Pacific, pp 1–8

Ekström J (2011) Mahalanobis' distance beyond normal distributions

Ernsberger D, Ikuesan AR, Venter HS, Zugenmaier A (2017) A web-based mouse dynamics visualization tool for user attribution in digital forensic readiness. In: 9th EAI International Conference on Digital Forensics & Cyber Crime, pp. 1–13

Fan XX, Chow KP, Xu F (2014) Web user profiling based on browsing behavior analysis. IFIP Adv Inf Commun Technol 433:57–71

Fielding RT et al (1999) RFC 2616: hypertext transfer protocol – HTTP/1.1

Guan X, Qin T, Member S, Li W, Wang P (2010) Dynamic feature analysis and measurement for large-scale network. Traffic Monitor 5(4):905–919

Hlavacs H, Kotsis G, Steinkellner C (1999) Traffic source modeling. Heport TR-99101. Inst Appl 1–12

Ikuesan AR, Venter HS (2018) Digital forensic readiness framework based on behavioral-biometrics for user attribution. In: 2017 IEEE Conference on Applications, Information and Network Security, AINS 2017, vol. 2018-Janua, no. 1, pp. 54–59

Ikuesan AR, Razak SA, Venter HS, Salleh M (2019) Polychronicity tendency-based online behavioral signature. Int J Mach Learn Cybern 10(8):2103–2118

Illankoon P, Tretten P, Kumar U (2019) Modelling human cognition of abnormal machine behaviour. Human-Intelligent Syst Integr 1(1):3–26

Iváncsy R, Juhász S (2007) Analysis of web user identification methods. World Acad Sci Eng Technol 34 2(3):338–345

Klema V, Laub A (1980) The singular value decomposition: its computation and some applications. IEEE Trans Automat Contr 25(2): 164–176

Lakhina A, Papagiannaki K, Crovella M, Diot C, Kolaczyk ED, Taft N (2004) Structural analysis of network traffic flows. ACM SIGMETRICS Perform Eval Rev 32(1):61

Lee JJJ, Gupta M (2007) A new traffic model for current user web browsing behavior. Intel Corp:1–7

Li B, Springer J, Bebis G, Gunes MH (2013a) A survey of network flow applications. J Netw Comput Appl 36(2):567–581

Li B, Springer J, Bebis G, Hadi Gunes M (Mar. 2013b) A survey of network flow applications. J Netw Comput Appl 36(2):567–581

Lin J, Williamson S, Borne KD, DeBarr D (2012) Pattern recognition in time series. In: Way MJ, Scargle JD, Ali KM (eds) Advances in Machine Learning and Data Mining for Astronomy, A. N. S. CRC Press, Taylor & Francis Group, pp 617–645

Mah B (1997) An empirical model of HTTP network traffic. In: Sixteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Driving the Information Revolution, IEEEI, NFOCOM'97, pp 592–600

Melnikov N, Schönwälder J (2010a) Cybermetrics: user identification through network flow analysis. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6155 LNCS, pp. 167–170,

Melnikov N, Schönwälder J (2010b) User identification based on the analysis of network flow patterns,

Mohlala M, Ikuesan AR, Venter HS (2018) User attribution based on keystroke dynamics in digital forensic readiness process. In: 2017 IEEE Conference on Applications, Information and Network Security, AINS 2017, vol. 2018-Janua, pp. 1–6

Mondal S, Bours P (2017) A study on continuous authentication using a combination of keystroke and mouse biometrics. Neurocomputing 230(November 2016):1–22

Murta C, Dutra G (2004) Modeling HTTP service times, Glob. Telecommun. Conf. …, pp. 972–976

Nguyen TD (2005) User behavior in web surfing user behavior in web surfing. National University of Singapore, Singapore

Padmanabhan B, Yang Y (2007) Clickprints on the web: are there signatures in web browsing data? Available SSRN http://ssrn.com/abstract=931057 or https://doi.org/10.2139/ssrn.931057,

Parthasarathy M (2005) A dissimilarity measure for comparing subsets of data: application to multivariate time series. In: Proceedings of the IEEE ICDM Workshop on Temporal Data Mining, pp 101–112

Pretorius S, Ikuesan AR, Venter HS (2018) Attributing users based on web browser history. In: 2017 IEEE Conference on Applications, Information and Network Security, AINS 2017, vol 2018-January, pp 1–6

Salleh M, Bakar A, Zaki A (2005) Comparison of TCP variants over self-similar traffic. In: In Proceedings of the Postgraduate Annual Research Seminar, pp 259–264

Schneider F, Ager B, Maier G, Feldmann A, Uhlig S (2012) Pitfalls in HTTP traffic measurements and analysis. In: Passive and Active Measurement, no. Section 2, Springer, pp. 242–251

Shaikh J, Fiedler M (2012) Modeling and analysis of web usage and experience based on link-level measurements. Teletraffic Congr. ITC …

Vedral V (2002) The role of relative entropy in quantum information theory. Rev Mod Phys 74(1):197–234

Velmurugan K, Mohamed MAM (2012) A study of network traffic pattern and its impact on performance in implementation of web services. Am … 5(1):63–69

Wang J, Hu X, Yang F, Luo H (2011) On modeling capacity and responsiveness of HTTP streaming. In: 2011 International Conference on Computer Science and Service System (CSSS). IEEE, pp 2358–2362

Woods J, Parish D, Phan R (2012) User and system identification using captured network traffic. In: 13th Annual Postgraduate symposium on the Convergence of Telecommunications, Networking and Broadcasting, pp 1–3

Yang YC (2010) Web user behavioral profiling for user identification. Decis Support Syst 49(3):261–271

Yang Y, Padmanabhan B (2010) Toward user patterns for online security: observation time and online user identification. Decis Support Syst 48(4):548–558

Zhang L (Apr. 2006) Thinking styles and the big five personality traits revisited. Pers Individ Dif 40(6):1177–1187

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.