Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2020.Doi Number

Mobile Social Service User Identification Framework Based on Action-Characteristic Data Retention

Chen-Yu Li[1,5], Chien-Cheng Huang[2], Feipei Lai[3], San-Liang Lee[4], Senior Member, IEEE, Jingshown Wu[1], Life Fellow, IEEE, Rong-Chi Chang[5], and Hsiang-Wei Huang[6]

[1]Graduate Institute of Communication Engineering and Department of Electrical Engineering, National Taiwan University, Taipei 10617, Taiwan, R.O.C.

[2]Department of Information Management, National Taiwan University, Taipei 10617, Taiwan, R.O.C.

[3]Computer Science and Information Engineering and Electrical Engineering Departments, National Taiwan University, Taipei 10617, Taiwan, R.O.C.

[4]Department of Electronic Engineering, National Taiwan University of Science and Technology, Taipei 10617, Taiwan, R.O.C.

[5]Department of Technology Crime Investigation, Taiwan Police College, Taipei 11696, Taiwan, R.O.C.

[6]Department of Computer Science, National Chengchi University, Taipei 11605, Taiwan, R.O.C.

Corresponding author: C.-Y. Li (e-mail: d00942021@ntu.edu.tw).

ABSTRACT Mobile social services are an indispensable part of our daily lives. These services are also favored by criminals because it is difficult to retrieve communication data from them. In the past, communication data provided by telecommunication carriers usually indicated when, from where, and with whom the communication occurred. Presently, it is difficult for law enforcement agencies and public security departments to obtain information regarding mobile social services. For this reason, these departments have requested Internet access service providers to store data that can be used to identify the user of a mobile social service at any given time. However, many non-government and civil society organizations claim that these practices violate privacy rights; hence, they strongly oppose the retention of the subscribers' data by the government. Currently, the European Union law does not allow "general and indiscriminate retention of traffic data and location data," except for "targeted" use against "serious crimes." Under this premise, ensuring the necessary data retention, while reducing the privacy violations and maintaining public security is a challenging task. In this study, a novel identification framework based on different types and action characteristics of mobile social services is proposed. Based on this framework, government agencies do not need to retain general and indiscriminate traffic data, but only data that aid in identification. Thus, this framework substantially reduces the volume of potential targets and improves the probability of correct target identification, ensuring a balance between privacy and public security.

INDEX TERMS data retention, Internet connection record, mobile social service, privacy, public security surveillance.

I. INTRODUCTION

The development of high-speed mobile Internet access and powerful mobile devices has led to the rapid growth of over-the-top (OTT) services. Mobile social services (MSSs) [1], a type of OTT service with connectivity and data sharing patterns among users, are the most popular among these applications and communication services. Their main functions include instant messaging and voice call services. One of the most famous providers, Facebook, had over 2.2 billion monthly active users and 1.45 billion daily active users as of May 2018 [2]. Facebook has more users than the population of China or India, two of the most populous countries in the world [3]. Approximately 100 billion messages are delivered and 3 billion minutes of voice and video calls are made from its applications and network per day [2]. In addition to Facebook, there are many similar MSSs worldwide, such as WhatsApp, Facebook messenger, WeChat, QQ, Instagram, Tumblr, Twitter, LINE, Skype, and Telegram [4]. Different MSSs are dominant in different countries [4].

In the past, voice call or message services were provided by local telecommunication carriers who maintained call detail records (CDRs) for billing purposes. When a carrier receives an authorization request from a local law

enforcement agency (LEA) or public safety department (PSD), it provides the CDRs or other communication data of the specific subscriber. These agencies can then apply this information in their works. MSS providers operate globally through the Internet, and have broken the barriers that limit service provision to local carriers. With the advantage of encrypted, secure, and free communication, people are increasingly using MSSs as their preferred communication tools. However, when a serious crime or an emergency occurs, LEAs and PSDs can no longer obtain information (metadata) regarding a specific communication or user from the MSS providers during the initial investigations, or immediately in the case of an emergency. Examples of some real scenarios in the daily operations of LEAs and PSDs include:

- A father with a violent criminal record threatened his wife that he would kill his daughter and then commit suicide. He had turned off his cellular service and used MSSs alone through an unknown Internet access connection to contact others.
- A businessman was kidnapped. The kidnapper used an MSS to contact the victim's family and demanded a ransom in Bitcoin.
- A student posted a suicide declaration on his social media site and then turned off his cellular phone.
- Defrauders posted fake product transaction information on their MSS site. They requested the victim to contact them privately for a larger discount. The victim believed them, contacted them, and then transferred the money. Finally, the defrauders disconnected from their contact and disappeared.

From these scenarios, governments understand that emergency services and criminal investigations require the support of MSSs. The main aims of government agencies in accessing Internet usage metadata are to identify the sender of a communication and the communication services that they use, in addition to determining whether a person has been accessing or creating available illegal material online [5]. Although a few MSS providers have legal request procedures for LEAs [6]–[10], they usually fail to fully meet the emergency requirements of global LEAs and PSDs owing to a lack of jurisdiction. Moreover, the response time [11], [12] is also a challenging issue in urgent cases.

In response to the abovementioned challenges, some countries have started to require local carriers and Internet access service providers (IASPs) to maintain records of metadata for their users' Internet access [13]–[16], such as the Internet connection record (ICR) in the UK. However, in real-world applications, there are many Internet links within a single carrier. Large volumes of traffic are transmitted along each link and are growing rapidly. This implies that each carrier produces significant amounts of ICRs each day, requiring a substantial network and storage equipment built into its core network. The carrier or government would require large budgets for this [17]–[19], and people have raised concerns regarding the cost and feasibility [18], [19]. Recently, the capability to capture ICRs has resulted in live trials in the UK [20], [21].

After Edward Snowden disclosed the National Security Agency (NSA) documents of the United States, people came to the realization that the government had been indiscriminately retaining their personal metadata from the Internet, thus violating their privacy [22]–[26]. The European Court of Justice (ECJ) declared that Directive 2006/24/EC (the Data Retention Directive) was invalid in 2014 [27], [28]. Nevertheless, the ECJ acknowledged that data retention is a valuable tool for governments in their pursuit of fighting serious crime and maintaining public security; however, the retaining standards need to be "appropriate" and "necessary" to realize their objectives [27], [28]. In 2016, the ECJ declared again that the general and indiscriminate retention of traffic and location data of all users that relate to all means of electronic communication is unlawful [29], [30]. This did not stop EU members from adopting legislation that permitted targeted retention of traffic and location data as a preventive measure. However, the legislation must carefully consider some conditions, such as the limit of data retention, categories of data to be retained, and retention period, are strictly necessary [29], [30]. Therefore, finding a balance between maintaining public security and privacy under restricted data retention policies is vital.

This study investigates the record of MSS actions during transactions through the Internet, with the objectives of determining if they can provide useful metadata, which metadata fields need to be retained, and the development of a suitable framework for the recording and analysis of ICRs. The main aim of this study is to find better methods for the rapid and reliable identification of an MSS subscriber that is of interest to LEAs and PSDs. The secondary aim is to achieve these methods within the scope of maintaining security and privacy while enabling unencumbered access for rescue or crime investigation purposes, and ensuring an efficient and cost-effective deployment.

II. RELATED WORK

When a serious crime that requires urgent attention is reported to a local PSD or LEA, they must respond immediately to enable prevention or rescue. The most critical issue is to determine who the suspect (or victim) is and where he/she is located. Typically, the Internet protocol (IP) address used by the specific MSS account is identified first. The IP address is then used to determine the possible terminal accounts from the IASP. These accounts are mapped to specific subscribers (i.e., the communication user's information) and the physical location (i.e., the served mobile communication base station or the WiFi hotspot location). However, it is challenging to obtain the IP address of a specific MSS user because the MSS provider and the IASP are usually different and independent. Additionally, it is difficult to request MSS providers to provide rapid assistance

outside of their located countries. Retaining Internet user metadata, which includes the IP address, port number, mobile phone number, and the service/domain at a specific time [31], seems to be a possible solution to obtain the IP address of a specific MSS user [5], [32]. Unfortunately, there are some critical issues in real-world applications. First, this method requires each telecommunication carrier to pre-retain all subscribers' metadata in its network and wait for the government to access it. Because most MSS users connect to their served hosts frequently [32]–[34], significant amounts of metadata are produced, which are required to be stored by carriers [32]–[34]. Furthermore, it is difficult for government agencies to directly use these metadata to identify an MSS user. Moreover, full retention not only increases the risk of privacy violation, but also significantly increases the cost [18], [19], [32], and almost all communications through MSSs are usually encrypted [35], [36]. These issues may result in the retention of invalid and redundant metadata [32].

A few studies have proposed methods for the identification of applications, user actions, and the operating system of a user by analyzing the recorded metadata [36]–[40]. These studies mainly focused on identifying user activities accurately. However, fewer studies have focused on the application of these methods to real LEA and PSD operations, such as identifying the sender of a communication, determining whether a person has been accessing or creating available illegal material online, or providing information to aid a rescue mission [5].

Lin et al. [41] discussed the application of a man-in-the-middle (MITM) based framework in a telecommunication carrier's network to defeat the communication of popular MSSs with secure socket layer/transport layer security (SSL/TLS). Their method required all users (whether targets or not) to pre-plug a custom or self-signed certificate into their certificate store [42], [43]. This induced critical impacts on the security and privacy. The certificate generated by a PSD or LEA will not have a valid trust chain, resulting in the application possibly terminating the connection instead of using a potentially insecure communication channel [42]. This in turn may result in the user being unable to connect to the MSS.

The proposed framework is based on the interactive actions of each MSS application. The MSS actions, such as sending a file, photograph, or video, or making a call, can be denoted in the record. The actions of each user can be sorted sequentially by the time of occurrence. This sequence serves as the fingerprint of the MSS user. There is an extremely low probability of having the same sequence matching for a long sequence for each user. There is a high probability that the user with the best-matched sequence is our target; thus, enabling a government agency to distinguish specific MSS users by their Internet access metadata. The MSS fingerprints can be obtained in several ways, such as from the chatting record of a communication party, through digital forensics, by requesting the metadata of the Internet, or through lawful interception.

The novelty of the proposed approach can be described as follows. (1) Owing to the characteristics of the MSS actions, IASPs do not need to pre-store metadata or only pre-store some metadata without storing ICRs for the entire network. (2) The MSS fingerprint mechanism is introduced to improve the identification accuracy of a potential target for the large volume of Internet access metadata. (3) The proposed approach can be applied to partial and fully encrypted MSS communications without breaking down the original security mechanism. (4) Identification can be performed across multiple MSSs.

III. FUNDAMENTAL EXPERIMENTS AND RESULTS

The most vital aspect of the proposed framework is that the government agency should be able to identify the characteristics of most MSS actions clearly and easily. For this purpose, a basic experiment is set up and its architecture is illustrated in Fig. 1. A rooted Samsung smartphone SM-G9208 (Android 7.0) with Android tcpdump (Version 4.9.3 / 1.9.1) [44] and two iPhone 6 Plus cell phones (iOS 12.4.5) with the remote virtual interface (RVI) mechanism [45] are mainly used for testing because the two main mobile operating systems, Android and iOS, possess more than 98% of the global market [46].

The MSSs tested are currently the most popular in the world. They included WhatsApp, Facebook Messenger, WeChat, Viber, Discord, Telegram, LINE, KaKaoTalk, Signal, imo, and VK [47]–[49]. Both Android and iOS editions exist for the above MSSs. Any significant difference in the traffic characteristics between the Android and iOS editions for the same MSS can be detected. In contrast, iMessage and FaceTime [50], which are pre-installed in the iPhone, provide a similar service as the above MSSs, and are tested with two identical iPhone 6 Plus cell phones in this study.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3009010, IEEE Access

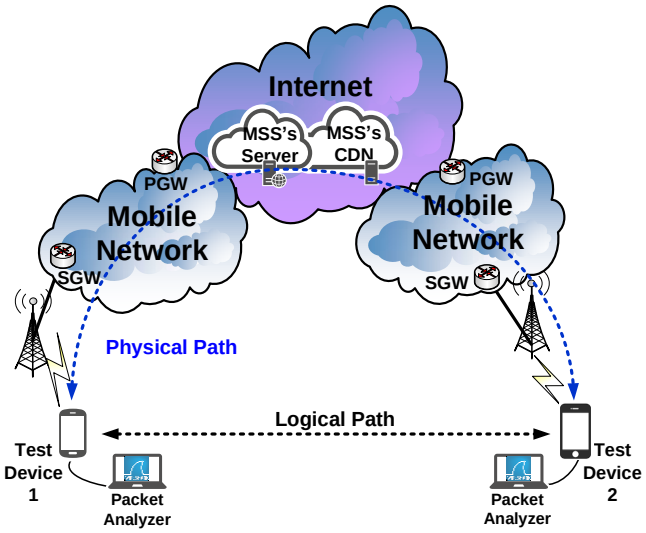


FIGURE 1. General observed architecture of MSS traffic.

The tested actions of each MSS on the sender and receiver sides were generated deliberately and sniffed the traffic from two tested devices, respectively. We examined the traffic characteristics of each action with a focus on the following points:

1) Is the traffic for the MSS encrypted?
2) Is there a common characteristic point in the traffic that can be found during the specific MSS used?
3) Is there a common characteristic point in the traffic that can be found during the specific action used?
4) Is it possible for the sending and receiving traffic to disclose the IP address of the two communication parties during a specific action?

More details of the tested actions and steps are presented in the following section.

A. TEST CONDITIONS

All tested information (e.g., supported actions, operating systems, and version) of the MSSs are listed in Table I. All actions of the MSSs, including sending and receiving different types of messages and making or answering voice/video calls, were investigated in this experiment. A few actions and the specific MSSs, such as sending a file through Facebook Messenger or all actions of iMessage, only support one operating system (e.g., iOS). Hence, two iPhone 6 plus cell phones are used such that have the same test conditions.

The MSSs can generate some traffic automatically, producing interference during the examination. For this purpose, on the sender side, we turned on the MSS under test, entered the conversation pane of the tested party, and left the device idle for a short time until we captured its traffic and tested it, thereby reducing the non-related traffic interference.

Each message-related action can be distinguished into three states: 1) the sender sends (SS), 2) the receiver clicks in the sender's conversation pane (RC), and 3) the receiver clicks the message to read (RR) (it only exists in the photo, voice, video, file, and location actions).

In addition, all call-related actions can also be distinguished in two states: 1) no-answer from the callee (NANS) and 2) callee answers (ANS). The traffic for all of

TABLE I
LIST OF TESTED MSS APPLICATIONS AND ACTIONS

Tested MSS	Supported operating system (Tested MSS version)	Tested actions
WhatsApp	Android (2.19.360) iOS (2.20.22)	1. Send and receive the text, sticker, photo, voice, video, file, and location message. 2. Dial and answer the voice/video call.
Facebook Messenger	Android (247.0.0.10.117) iOS (252.1)	1. Send and receive the text, sticker, photo, voice, video, file (only in iOS edition), and location message. 2. Dial and answer the voice/video call.
WeChat	Android (7.0.10) iOS (7.0.10)	1. Send and receive the text, sticker, photo, voice, video, file, and location message. 2. Dial and answer the voice/video call.
Viber	Android (12.2.0.7) iOS (12.3.5)	1. Send and receive the text, sticker, photo, voice, video, file, and location message. 2. Dial and answer the voice/video call.
Discord	Android (10.4.2) iOS (3.2.0)	1. Send and receive the text, sticker, photo, and file message. 2. Dial and answer the voice/video call.
Telegram	Android (5.14.0) iOS (5.15)	1. Send and receive the text, sticker, photo, voice, video, file, and location message. 2. Dial and answer the voice call.
LINE	Android (10.2.1) iOS (10.2.0)	1. Send and receive the text, sticker, photo, voice, video, file, and location message. 2. Dial and answer the voice/video call.
KaKaoTalk	Android (8.7.7) iOS (8.7.6)	1. Send and receive the text, sticker, photo, voice, video, file, and location message. 2. Dial and answer the voice/video call.
Signal	Android (4.55.8) iOS (3.5.0)	1. Send and receive the text, sticker, photo, voice, video, file, and location message. 2. Dial and answer the voice/video call.
imo	Android (2020.2.51) iOS (2020.2.2)	1. Send and receive the text, sticker, photo, voice, video, and file message. 2. Dial and answer the voice/video call.
VK	Android (5.54) iOS (5.34.2)	1. Send and receive the text, sticker, photo, voice, video, file, and location message. 2. Dial and answer the voice/video call.
iMessage	Embedded in iOS	Send and receive the text, sticker, photo, voice, video, file, and location message.
FaceTime	Embedded in iOS	Dial and answer the voice/video call.

the above states was recorded and examined individually. The tested steps of each action are as follows:

Step 1: Select one of the devices as the sender (the caller in a call action), and the other as the receiver (the callee in a call action).

Step 2: Enter the conversation pane of the MSS under test on the sender's device. Stand still for a short period of time and keep all test devices idle. Then, capture the traffic on both sides.

Step 3-1 (For the test message-related actions): [SS state] Perform the test action on the sender's device and wait until this action is successfully received on the receiver's device. Stop to capture the traffic and store it.

Step 3-2 (For the test message-related actions): [RC state] Start to capture the sender's traffic for the receiver side again. The receiver clicks in the sender's conversation pane. Stop to capture the traffic and store it.

Step 3-3 (For the test message-related actions): [RR state] Start to capture the traffic for the sender and the receiver sides again. The receiver clicks the message to read it. Stop to capture the traffic and store it. (This step is for photo, voice, video, file, and location actions.)

Step 4-1 (For the test call-related actions): [NANS state] Perform the call action on the caller's device and wait until this action is successfully received on the callee's device. The callee does not answer the call action during the call ring period end. Then, stop to capture the traffic and store it.

Step 4-2 (For the test call-related actions): [ANS state] Perform the call action on the caller's device and wait until this action is successfully received on the callee's device. The callee answers the call action during the call ring period. Keep this conversation going for a short period of time and then hang up. Then, stop to capture the traffic and store it.

Step 5: Exchange the sender and receiver roles and then repeat steps 2–4.

Step 6: End the test of this action until all roles are played by the test devices.

The workflow of the test steps is illustrated in Fig. 2. The test results are presented in the next section.

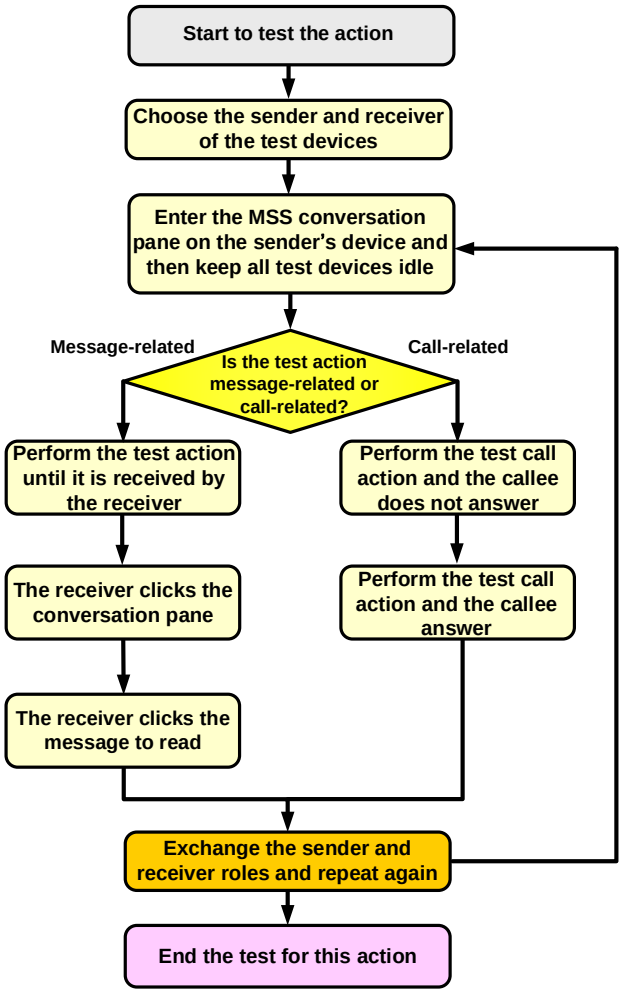


FIGURE 2. Workflow of the test MSS actions.

B. GENERAL OBSERVATIONS

All actions listed in Table I are tested and observed, and are summarized in Table II. Although most of the traffic for the MSSs are encrypted, some characteristics can be observed. These characteristics clearly indicate some types of MSSs or the actions that may have been performed. For example, the domain name requests can be used to identify what the MSS is used for (Table III).

One action can generate multiple DNS requests and more than one action generates the same DNS requests. Therefore, it is often difficult to know the name of the DNS request/response that corresponds to the specific action of the MSS. Thus, more characteristics need to be considered to identify the action the user performed.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3009010, IEEE Access

TABLE II
SUMMARY OF COMMON SESSION CHARACTERISTICS FOR POPULAR MSSS AND THEIR ACTIONS

Tested MSS	Is the traffic encrypted?	Is there at least one piece of DNS request information that can be related to the MSS?	Is there at least one piece of DNS request information that can correspond to the action of the MSS during the actions used?	Is there at least one session that is related to the MSS? A large volume or other clear characteristics of the traffic for this action are used	Is it possible for at least one UDP session to be related to the MSS for the actions performed?	Is it possible for at least one session to discloses the IP address of the communication parties during the actions used?
WhatsApp	All actions	Yes	Not all (Multiple actions correspond to the common requests)	Except the text action	Possible in voice and video call actions	Possible in voice and video call actions
Facebook Messenger	All actions	Yes	Not all (Multiple actions correspond to the common requests)	Except the text action	Possible in voice and video call actions	Possible in voice and video call actions
WeChat	All actions	Yes	No	Except the text action	Possible in voice and video call actions	Possible in voice and video call actions
Viber	All actions	Yes	Not all (Multiple actions correspond to the common requests)	Except the text action	Possible in voice and video call actions	Possible in voice and video call actions
Discord	All actions	Yes	Not all (Multiple actions correspond to the common requests)	Except the text and sticker action	Possible in voice and video call actions	None
Telegram	All actions	No	No	Except the text action	Possible in voice call action	Possible in voice call action
LINE	All actions	Yes	Not all (Multiple actions correspond to the common requests)	Except the text action	Possible in voice and video call actions	Possible in voice and video call actions
KaKaoTalk	Not All (Except the voice action)	Yes	Not all (Multiple actions correspond to the common requests)	Except the text action	Possible in voice and video call actions	None
Signal	All actions	Yes	Not all (Multiple actions correspond to the common requests)	Except the text action	Possible in voice and video call actions	Possible in voice and video call actions
imo	Not All (Except the sticker action)	Yes	Not all (Multiple actions correspond to the common requests)	Except the text action	Possible in file (sending), voice and video call actions	Possible in voice and video call actions
VK	All actions	Yes	Not all (Multiple actions correspond to the common requests)	Except the text action	Possible in voice and video call actions	Possible in voice and video call actions
iMessage	All actions	Yes	Not all (Multiple actions correspond to the common requests)	Except the text action	No	No
FaceTime	All actions	Yes	Not all (Multiple actions correspond to the common requests)	All actions	Possible in voice and video call actions	Possible in voice and video call actions

The volume (in packets or bytes) of the traffic is usually larger than the text message action when the users send/receive the sticker, photograph, voice, video, file, or location. There is at least one session in which the amounts of the user's outgoing/incoming packets or bytes during the photo, voice, video, and file actions have significant differences and are non-symmetric. This can be used to identify the user's role. One possible role is that of a sender when the outgoing traffic is larger than the incoming session. The other possible role is that of a receiver when the incoming traffic is larger than the outgoing session. Additionally, the magnitude of the sender's outgoing traffic is approximate to the magnitude of the receiver's incoming traffic in the above sessions. We can calculate the difference using the following formula.

$$\Delta_{packets\ or\ bytes}\% = \frac{|N_{In,Receiver} - N_{Out,Sender}|}{N_{Out,Sender}} \times 100\% \quad (1)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3009010, IEEE Access

TABLE III
RELATIONSHIPS BETWEEN USED MSSS AND REQUESTED DOMAIN NAMES

Used MSS	Requested domain names for the MSS used
WhatsApp	*.whatsapp.net
Facebook Messenger	*.facebook.com *.fbcdn.net *.fbsbx.com
WeChat	*.qq.com *.tencent-cloud.net
Viber	*.viber.com
Discord	discordapp.com *.discordapp.com *.discord.media gateway.discord.gg media.discordapp.net
LINE	lan.line.me *.line.naver.jp *.line-scdn.net nelo2-col.linecorp.com obs-tw.line-apps.com
KaKaoTalk	*.kakao.com *. kakaocdn.net
Signal	textsecure-service.whispersystems.org cdn.signal.org turnX.whispersystems.org
imo	imo.im api.imotech.tech *.imoim.app
VK	vk.com *.vk.com *.vk-cdn.net *.vk-portal.net *.userapi.com clientapi.mail.ru stun.mail.ru
iMessage and FaceTime	apple.com *.apple.com *.apple-dns.net *.icloud.com *.icloud-content.com

Here, $N_{In,Receiver}$ and $N_{Out,Sender}$ represent the magnitude of the receiver's incoming traffic (in packets or bytes) and the sender's outgoing traffic, respectively.

The traffic differences in the photo, voice, video, and file actions for each tested MSS are listed in Table IV. For most MSSs, there exists a session in which the traffic between the sender and receiver is similar during the above-mentioned actions, particularly in bytes. Although the traffic between the sender and receiver is not similar in a few cases, we find through multiple tests that the $\Delta_{packets\ or\ bytes}$ % is similar in these cases, which is also a characteristic.

We can observe the IP addresses of the connected hosts in the above sessions, which usually correspond to the name of the specific requested domain or the WHOIS information of the IP addresses for the MSS (Table V). In addition, the state in which these sessions are generated should be considered.

The above-mentioned sessions on the sender side were generated in the SS state. However, on the receiver side, not all of the sessions for the above actions were generated in the SS state, particularly in the video and file actions (Table VI). This implies that the other states (RC and RR states) are the key states that trigger the main session. This is useful for government agencies in developing their identification strategies, and can be used to distinguish the types of actions that are used on the receiver side when the volume of the users' traffic is close, in addition to providing information regarding whether the receiver has read the message or not.

In the location action, there is no clear traffic characteristic that is comparable with the above message-related actions. The domain name request is a crucial characteristic that can identify the user's action. We determined that the users connect to the MSS servers during the location actions and the specific domain names were requested (Table VII). Some of the listed domain names were also commonly used in the different MSSs. Additionally, the main traffic of these users usually corresponds to these domain names.

Next, we consider the call-related actions. When the users establishes a voice or video call using the MSS, the caller connects to the callee instantly and then the callee's phone rings. We identified that the common characteristics of the call actions are different from those of other actions (Table VIII). First, there is at least one session that is used for the user datagram protocol (UDP) during these call actions for most MSSs. The caller and callee usually use this protocol to connect the servers or the call party. When they connect to the MSS servers, there are usually some characteristics, such as the domain name, resolved name, WHOIS, or an autonomous system number (ASN) of the servers, that are related to the MSS, or they use the specific port number (Table IX).

Second, the IP addresses of the call parties are usually disclosed in the ANS state and the UDP sessions used. In some MSSs, the IP addresses of the call parties are possibly disclosed in the NANS state.

Sometimes, when both call parties connect to the servers, their IP addresses are not disclosed in some MSSs (or in the NANS state). However, we find that both call parties connect to the servers not only with the characteristics in Table IX (the same domain name request, the IP addresses of the server, or the port number) but also the difference between the duration of the both call parties' mainly sessions is small (Table X). These characteristics imply that the call parties can be correlated although their IP addresses are not disclosed.

All of the above call action characteristics are excellent, and can be used to identify a caller and callee faster than the other actions. For this reason, it is necessary to determine the conditions that can activate these call actions successfully. These conditions are listed in Table XI. We find that many MSSs can establish a call action successfully although the caller knows only the callee number. This implies that they can be used to instantly generate the characteristics of the call action on the callee side, which is suitable for application in some urgent cases.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3009010, IEEE Access

TABLE IV
DIFFERENCE BETWEEN THE MAGNITUDE OF OUTGOING AND INCOMING TRAFFIC DURING PHOTO, VOICE, VIDEO, AND FILE ACTIONS

Tested actions		Photo		Voice		Video		File	
Tested MSSs		Android sends	iPhone sends	Android sends	iPhone sends	Android sends	iPhone sends	Android sends	iPhone sends
WhatsApp	$\Delta_{packets}$%	8.92	5.80	8.92	1.92	10.52	7.36	8.88	10.67
	Δ_{bytes}%	2.11	1.94	2.11	1.15	1.16	2.86	1.10	0.43
Facebook Messenger	$\Delta_{packets}$%	16.16	0.28	4.45	10.58	6.23	16.74	0.08*	10.83
	Δ_{bytes}%	11.00	10.89	0.77	1.22	12.20	0.47	0.37*	0.52
WeChat	$\Delta_{packets}$%	11.96	0.35	34.45	18.91	7.91	16.44	10.76	15.67
	Δ_{bytes}%	2.08	2.28	9.88	4.47	0.89	1.78	1.28	1.78
Viber	$\Delta_{packets}$%	10.41	2.10	10.3	4.76	9.63	13.04	9.67	13.13
	Δ_{bytes}%	1.97	1.41	1.05	1.58	0.22	2.57	0.18	2.67
Discord	$\Delta_{packets}$%	12.99	8.66	No this action	No this action	No this action	No this action	12.66	11.12
	Δ_{bytes}%	2.04	0.47	No this action	No this action	No this action	No this action	1.61	0.04
Telegram	$\Delta_{packets}$%	3.34	7.96	5.12	3.88	2.01	3.53	19.55	6.22
	Δ_{bytes}%	3.97	4.81	3.90	0.17	4.25	7.75	22.60	1.72
LINE	$\Delta_{packets}$%	12.75	8.07	11.86	5.88	11.41	13.27	10.71	15.51
	Δ_{bytes}%	0.76	1.95	0.28	2.57	1.76	1.45	1.32	1.58
KaKaoTalk	$\Delta_{packets}$%	14.28	6.38	273.86	15.84	56.40	11.78	11.13	11.70
	Δ_{bytes}%	2.30	1.85	321.97	5.01	51.75	1.20	1.50	1.30
Signal	$\Delta_{packets}$%	1.78	30.02	2.18	28.84	0.41	30.84	1.10	30.68
	Δ_{bytes}%	0.58	3.46	0.76	3.26	0.35	3.35	0.59	3.34
imo	$\Delta_{packets}$%	46.15	27.10	7.62	25.48	45.57	27.16	49.79	15.55
	Δ_{bytes}%	18.03	11.38	1.22	4.90	7.46	5.91	5.00	0.15
VK	$\Delta_{packets}$%	2.87	12.12	9.67	10.00	28.42	3.10	8.87	7.03
	Δ_{bytes}%	10.20	6.41	1.50	4.09	29.03	5.65	0.03	1.91
iMessage*	$\Delta_{packets}$%	—	35.96	—	20.00	—	25.09	—	21.58
	Δ_{bytes}%	—	19.86	—	1.87	—	4.89	—	0.19

*(This action or only the MSS in the iOS edition; two iPhones were used to test this item.)

TABLE V
INFORMATION OF CONNECTED HOSTS RELATED TO MAIN SESSIONS DURING PHOTO, VOICE, VIDEO, AND FILE ACTIONS

Tested actions / Tested MSSs	Photo	Voice	Video	File
WhatsApp	mmg.whatsapp.net (or mmx-ds.cdn.whatsapp.net) media.ftpeX-X.fna.whatsapp.net (only in the receiver)	mmg.whatsapp.net (or mmx-ds.cdn.whatsapp.net)	mmg.whatsapp.net (or mmx-ds.cdn.whatsapp.net) media.ftpeX-X.fna.whatsapp.net	mmg.whatsapp.net (or mmx-ds.cdn.whatsapp.net)
Facebook Messenger	rupload.facebook.com (or star.c10r.facebook.com) scontent.xx.fbcdn.net scontent.ftpeX-X.fna.fbcdn.net (only in receiver)	rupload.facebook.com (or star.c10r.facebook.com) cdn.fbsbx.com (or scontent.xx.fbcdn.net)	rupload.facebook.com (or star.c10r.facebook.com) video.xx.fbcdn.net	rupload.facebook.com (or star.c10r.facebook.com) cdn.fbsbx.com (or scontent.xx.fbcdn.net)
WeChat	IP range: 203.205.X.X (The WHOIS information can be related to WeChat.)			
Viber	media-share-8.s3.ap-northeast-1.amazonaws.com (or s3-r-w.ap-northeast-1.amazonaws.com) dl-media.viber.com (or d1fje9gm3d05t8.cloudfront.net)			
Discord	discordapp.com (or cdn.discordapp.com) media.discordapp.net	(No this action)	(No this action)	discordapp.com (or cdn.discordapp.com)
Telegram	IP range: 91.108.X.X (The WHOIS information can be related to Telegram.)			
LINE	obs-tw.line-apps.com (or obs-jp-addr.line-apps.com)			
KaKaoTalk	The IPs of the connected hosts is variable, and there is no clear range. (The ASN of the IP address ranges are related to the Kakao Corp.)	up-a.talk.kakao.com (or up-a.talk.gl.kakao.com) dn-a2.talk.kakao.com (or dn-a.talk.gl.kakao.com)	The IPs of the connected hosts are variable, and there is no clear range. (The ASN of the IP address ranges are related to the Kakao Corp.)	The IPs of the connected hosts are variable, and there is no clear range. (The ASN of the IP address ranges are related to the Kakao Corp.)
Signal	cdn.signal.org (or d83eunklitikj.cloudfront.net)			
imo	IP range 104.36.224.X (The ASN of the IP address ranges are related to imo.)			
VK	sunX-XX.userapi.com puX-XX.vk-cdn.net	pu.userapi.com puX-XX.vk-cdn.net psvX.userapi.com (or ps.userapi.com)	sunX-XX.userapi.com vu.vk.com (or pu.vk.com)	sunX-XX.userapi.com puX-XX.vk-cdn.net psvX.userapi.com (or ps.userapi.com)
iMessage	edge-XXX.hkhkg.ce.apple-dns.net gateway-asset.ce.apple-dns.net			

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3009010, IEEE Access

TABLE VI
STATE OF RECEIVER CONNECTED MSS HOSTS RELATED TO MAIN SESSIONS DURING PHOTO, VOICE, VIDEO, AND FILE ACTIONS

Tested actions / Tested MSSs	Photo	Voice	Video	File
WhatsApp	SS	SS	RR (Generates traffic of the thumbnail in the SS state.)	RR
Facebook Messenger	RR	SS	RR (Generates traffic of the thumbnail in the SS state.)	RR
WeChat	RR (Generates traffic of the thumbnail in other states.)	SS (in Android receives) RC (in iPhone receives)	RR (Generates traffic of the thumbnail in the RC state.)	RR
Viber	SS	SS	RR (Generates traffic of the thumbnail in the SS state.)	RR
Discord	RR (Generates traffic of the thumbnail in the RC state.)	(No this action)	(No this action)	RR
Telegram	SS	RC	RR (received in Android) RC (received in iPhone)	RR
LINE	RR (received in Android) SS (received in iPhone)	SS	RC (Generates traffic of the thumbnail in the SS state.)	RC (received in Android) RR (received in iPhone)
KaKaoTalk	SS	RR	RR (Generates traffic of the thumbnail in the RC state.)	RR
Signal	SS	SS	RR	RR (received in Android) SS (received in iPhone)
imo	SS	SS (received in Android) RC (received in iPhone)	RR	RR
VK	SS	SS (received in Android) RC (received in iPhone)	RR (Generates traffic of the thumbnail in the RC state.)	RC (received in Android) RR (received in iPhone)
iMessage	SS	SS	SS	RR

TABLE VII
LIST OF SPECIFIC DOMAIN NAME REQUESTS DURING THE LOCATION ACTION

Requested domain name	MSSs
clients4.google.com (or client.l.google.com)	WhatsApp, Telegram, LINE, KaKaoTalk, Signal, VK, iMessage
csi.gstatic.com	WhatsApp, Viber, Telegram, LINE, KaKaoTalk, Signal, VK
gspeXX-ssl.ls.apple.com gsp-ssl.ls.apple.com	WhatsApp (in iPhone), Facebook Messenger, WeChat, Viber, Telegram, LINE, KaKaoTalk, Signal, VK
gs-loc.apple.com	WeChat (in iPhone)
maps.googleapis.com	WhatsApp, WeChat, Viber, KaKaoTalk
semanticlocation-pa.googleapis.com	Facebook Messenger
*.map.qq.com p0.map.gtimg.com	WeChat
maps.google.com	WeChat, iMessage
map-ce.viber.com	Viber
firebaseremoteconfig.googleapis.com	Viber
maps.gstatic.com	Viber
geomobileservices-pa.googleapis.com	Viber, Telegram, LINE, Signal
maps.app.goo.gl	Viber
dmaps.daum.net ot1.maps.daum-img.net	KaKaoTalk
footprints-pa.googleapis.com	VK
firebasedynamiclinks-ipv4.googleapis.com firebasedynamiclinks-ipv6.googleapis.com	iMessage

C. SUMMARY

This study found that, although the traffic is encrypted, some characteristics can be used to identify the MSSs used, the actions, and the communication party of the user. The most crucial characteristics that need to be retained include the requested domain name, IP addresses of connected servers, volume of traffic, transport layer protocol (TCP or UDP), port number, duration of characteristic sessions, and disclosed IP addresses of communication parties.

These characteristics cover a majority of the commonly used actions and MSSs. Even if a few individual actions of the MSS do not possess common characteristics, other characteristics can be determined; these characteristics can also be retained. For instance, there is no clear range of the IP addresses for connected servers in the imo's call actions (Table IX). However, the action includes a TCP session that connects to the MSS server and also accompanies a UDP session with a high port number, small traffic, and a long duration. This is a unique characteristic that can be used to identify the call action of the MSS. This coincides with the IP addresses and the characteristics for possible users.

Based on the results obtained via the abovementioned experiments, a novel IP data retention framework capable of evaluating the identification of a specific MSS's user is proposed. The operation details and evaluation results are presented in subsequent sections.

TABLE VIII
SUMMARY OF COMMON TRAFFIC CHARACTERISTICS DURING VOICE OR VIDEO CALL ACTIONS

Characteristics State of call actions	There is at least one session that uses the UDP to communicate	IP addresses of the call parties are disclosed	There is at least one session where both call parties are connected with the same domain name, IP address, or the port number of the servers	There is at least one session where the difference between the duration of both call parties is small
Callee does not answer (The NANS state)	WhatsApp, Facebook Messenger, WeChat(only in Android callee), Viber, Discord (only in caller), LINE, KaKaoTalk, Signal, imo, VK	Facebook Messenger, Signal, imo, VK (only in iPhone caller)	WhatsApp, Facebook Messenger, WeChat, Viber, Telegram*, LINE, KaKaoTalk, Signal, imo, VK, FaceTime	WhatsApp, Facebook Messenger, WeChat (only in Android callee), Viber, LINE, KaKaoTalk, Signal, imo, VK (only in iPhone caller), FaceTime
Callee answers (The ANS state)	WhatsApp, Facebook Messenger, WeChat, Viber, Discord, Telegram*, LINE, KaKaoTalk, Signal, imo, VK, FaceTime	WhatsApp, Facebook Messenger, WeChat, Viber, Telegram*, LINE, Signal, imo, VK, FaceTime	WhatsApp, Facebook Messenger, WeChat, Viber, Discord, Telegram*, LINE, KaKaoTalk, Signal, imo, VK, FaceTime	WhatsApp, Facebook Messenger, WeChat, Viber, Telegram*, LINE, KaKaoTalk, Signal, imo, VK, FaceTime

(*: Telegram does not provide the video call action.)

TABLE IX
MAJOR CHARACTERISTICS FOR SERVERS WHERE BOTH CALL PARTIES ARE CONNECTED DURING CALL ACTIONS

Tested MSS	Domain name or possible IP range of connected servers that are related to the MSS	The used protocol	The used port number
WhatsApp	edgeray-shv-XX-XXXX.facebook.com	UDP	3478
Facebook Messenger	edge-stun.facebook.com	UDP	3478
	edge-turnservice-shv-XX-XXXX.facebook.com	UDP	Variable (five digits)
WeChat	IP range: 124.156.X.X and 203.205.X.X (Its WHOIS information can be related to this MSS.)	UDP (and TCP)	80, 8000, 8080, or 16285
Viber	IP range: 54.64.191.X (Its WHOIS information can be related to this MSS.)	UDP	7985 or 7987
Discord	hongkongXXX.discord.media	TCP	Variable (five digits)
	IP range: 43.239.137.X (The ASN of the connected IP address is AS49544 [i3D.net B.V].)	UDP	Variable (five digits)
Telegram	IP range: 91.108.X.X (Its WHOIS information can be related to Telegram Messenger Network.)	UDP (or TCP)	Variable (three digits)
LINE	IP range: 147.92.130.X (Its WHOIS information can be related to this MSS.)	UDP	Variable (five digits)
KaKaoTalk	IP range: 139.150.5.X and 211.231.105.X The ASNs of the connected IP address are AS10158 and AS38099 [Kakao Corp].)	UDP	Variable (five digits)
Signal	turn1.whispersystems.org	UDP	80 or 443
imo	IP range: No clear range (The ASNs of the servers' IP addresses are also not available. They include Google, AOFEI Data International Company Limited, T.H. Global Vision SARL, Locknet, and others.)	UDP	Variable (four-five digits)
VK	calls.vk.com	TCP	80
	stun.mail.ru	UDP	3478
	IP range: 95.213.27.X (Its WHOIS information can be related to this MSS.)	UDP	443, 3478, or variable (five digits)
FaceTime	XX.courier-push-apple.com.akadns.net	TCP	5223

IV. PROPOSED FRAMEWORK AND EVALUATION RESULTS

The previous section discussed the characteristics of Internet traffic at the sender and receiver sides for several popular MSSs. When a person has sent/received data to/from someone through specific actions and MSSs, corresponding characteristics for both users exist. This implies that it is possible for agencies to pre-record ICRs with specific characteristics of MSSs. These data can then be used to identify the suspects. Based on these results, a novel framework is proposed and the performance of this framework is evaluated.

A. ARCHITECTURE OF THE FRAMEWORK

Characterization of Internet traffic is applied for the management and supervision of telecommunication carriers and IASPs [51], [52]. For this purpose, telecommunication carriers or IASPs collect traffic information using technologies such as NetFlow [52]. The proposed framework is depicted in Fig. 3; it employs equipment and a configuration similar to that used to collect and analyze traffic information.

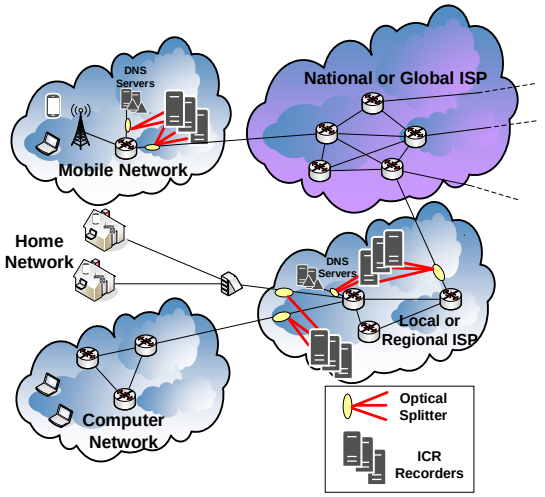


FIGURE 3. General framework applied to ICR recorders.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3009010, IEEE Access

TABLE X
DURATION AND ITS DIFFERENCE IN MAIN SESSIONS OF CALLER AND CALLEE DURING CALL ACTIONS

Tested MSS (Unit: s)	Type of call actions	Test state	Connect to server or call party	Duration of the caller (Use Android) (A)	Duration of the callee (Use iPhone) (B)	Difference |A-B|	Duration of the caller (Use iPhone) (C)	Duration of the callee (Use Android) (D)	Difference |C-D|
WhatsApp	Voice	No answer	Server	51.1	45.0	6.1	43.9	43.3	0.6
		Answer	Call party	61.4	61.4	0.0	60.8	61.2	0.4
	Video	No answer	Server	50.3	45.2	5.1	44.2	43.1	1.1
		Answer	Call party	65.1	65.1	0.0	60.0	60.2	0.2
Facebook Messenger	Voice	No answer	Call party	59.7	59.6	0.1	57.9	57.9	0.0
		Answer	Call party	81.0	81.1	0.1	72.3	72.5	0.2
	Video	No answer	Call party	58.6	59.4	0.8	58.7	58.7	0.0
		Answer	Call party	73.9	73.7	0.2	76.8	76.7	0.1
WeChat	Voice	No answer	Server	56.5	58.0	1.5	58.0	57.4	0.6
		Answer	Server	81.9	82.1	0.2	82.6	82.2	0.4
	Video	No answer	Server	30.4	30.1	0.3	58.0	56.5	1.5
		Answer	Server	80.7	80.5	0.2	77.0	76.7	0.3
Viber	Voice	No answer	Server	40.0	37.6	2.4	39.8	38.0	1.8
		Answer	Call party	61.7	61.7	0.0	60.7	60.7	0.0
	Video	No answer	Server	40.0	37.2	2.8	39.8	38.3	1.5
		Answer	Call party	69.2	70.2	1.0	65.7	65.8	0.1
Telegram	Voice	Answer	Call party	61.9	61.9	0.0	62.9	62.9	0.0
LINE	Voice	No answer	Server	60.4	58.4	2.0	60.4	58.2	2.2
		Answer	Call party	46.1	46.1	0.0	45.5	45.5	0.0
	Video	No answer	Server	60.4	56.6	3.8	60.5	59.8	0.7
		Answer	Call party	53.1	57.4	4.3	48.4	48.4	0.0
KaKaoTalk	Voice	No answer	Server	5.0	5.2	0.2	5.0	5.0	0.0
		Answer	Server	61.9	61.1	0.8	60.7	60.6	0.1
	Video	No answer	Server	5.0	5.2	0.2	5.0	5.8	0.8
		Answer	Server	77.2	73.5	3.7	61.5	61.6	0.1
Signal	Voice	No answer	Call party	118.0	118.0	0.0	118.6	118.5	0.1
		Answer	Call party	84.4	84.4	0.0	78.5	78.5	0.0
	Video	No answer	Call party	118.7	118.7	0.0	116.3	116.2	0.1
		Answer	Call party	85.7	85.7	0.0	104.4	104.3	0.1
imo	Voice	No answer	Call party	60.8	60.9	0.1	69.5	69.6	0.1
		Answer	Call party	103.0	102.9	0.1	77.5	77.5	0.0
	Video	No answer	Call party	60.4	60.7	0.3	69.4	69.6	0.2
		Answer	Call party	74.5	74.2	0.3	77.0	77.1	0.1
VK	Voice	No answer	Depends on the used device	59.6	58.7	0.9	47.9	48.0	0.1
		Answer	Call party	61.7	61.7	0.0	63.1	63.2	0.1
	Video	No answer	Depends on the used device	59.6	58.8	0.8	47.5	47.8	0.3
		Answer	Call party	62.8	62.8	0.0	62.7	62.7	0.0
FaceTime*	Voice	No answer (Use TCP)	Server	37.0	34.6	2.4	*(As the same two iPhones are used to test this MSS, these values are identical to those on the left.)		
		Answer	Call party	81.5	81.4	0.1			
	Video	No answer (Use TCP)	Server	36.9	34.5	2.4			
		Answer	Call party	80.1	80.1	0.0			

TABLE XI
CONDITIONS FOR VOICE/VIDEO CALLS

Condition of the MSS for voice/video calls	Tested MSS.
Caller needs to know the number of the callee before making the call.	WhatsApp, Viber, Telegram, KaKaoTalk, Signal, imo, and FaceTime.
Callee needs to add the caller to the contact list of the MSS before making the call.	Facebook Messenger, WeChat, Discord, LINE, and VK.

Optical splitters and ICR recorders are deployed on the backbones of telecommunication carriers and IASPs. Their function is to duplicate and record network traffic information. The proposed ICR recorders mainly collect traffic information that can be used to identify specific user actions of different MSSs. They are similar to the ICR system introduced in the UK [53]. In addition to recording the source and destination IP addresses and ports, they also record useful information such as the protocol (TCP or UDP), amount of traffic (bytes and packets), and conversation duration. The DNS query and response are also important characteristics that can be used to identify a MSS and the actions used. The entire conversations of each user are not recorded; only the conversations related to identified MSSs are recorded. This can reduce the impact

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3009010, IEEE Access

on privacy as well as the deployment costs. Examples of the proposed ICRs are shown in Fig. 4 and 5.

Many people use popular MSSs and the services produce significant volumes of ICRs and DNS queries/responses simultaneously in the networks of each telecommunication carrier and IASP. Fortunately, in real-world applications, government agencies have usually received MSS chatting records from one of the communicating parties, such as a reporter or the victim. These records can be analyzed to find the actions and their time of occurrence. Since the ICRs with specific action-characteristics are pre-stored by the telecommunication carriers and IASPs, government agencies can request the ICRs for a narrowed list of possible users. For this purpose, we developed two strategies and the details of the strategies are described in the next section.

B. APPLIED STRATEGIES FOR RETAINED ICRs

When considering a national level network, the deployment of ICR recorders for hundreds or thousands of backbones, and numerous MSS users, significant volumes of ICRs are continuously generated. The results from this study can be used to reduce the range of possible targets from the large volume of ICRs. However, there are several user send/receive actions featuring the same characteristic traffic simultaneously. Currently, it is difficult to precisely identify the user. Therefore, elucidating approaches to apply these results in real-world applications is essential. For this purpose, two applied strategies are proposed:

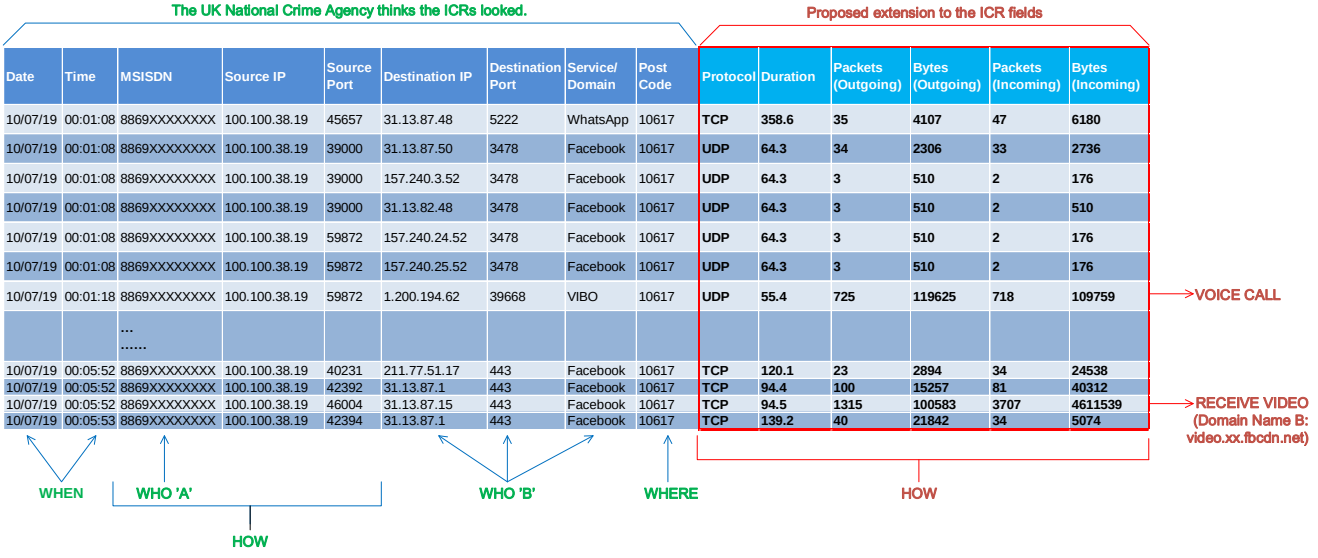
1) POST-ACTION SEQUENCE MAPPING

For this strategy, government agencies should request telecommunication carriers and IASPs to pre-record ICRs with some characteristics of the MSSs' actions, such as the characteristics discussed in the previous section. After government agencies have received MSS chatting records or other information provided from victims or reporters, they follow the workflow shown in Fig. 6. The steps included in this workflow are as follows:

Step 1: Attempt to determine identifiable pre-recorded actions that exist in received MSS records.

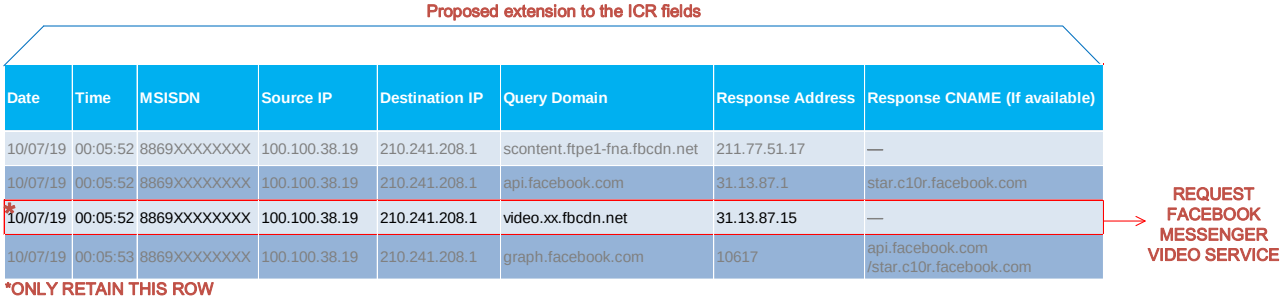
Step 2: If identifiable actions are found, each action with its occurrence time will be denoted as a sequence. Fig. 7 presents an example.

Step 3: Based on the sequence order, the government agency requests the pre-recorded ICRs of the specific MSS from the telecommunication carriers and IASPs.



Date	Time	MSISDN	Source IP	Source Port	Destination IP	Destination Port	Service/ Domain	Post Code	Protocol	Duration	Packets (Outgoing)	Bytes (Outgoing)	Packets (Incoming)	Bytes (Incoming)
10/07/19	00:01:08	8869XXXXXXXX	100.100.38.19	45657	31.13.87.48	5222	WhatsApp	10617	TCP	358.6	35	4107	47	6180
10/07/19	00:01:08	8869XXXXXXXX	100.100.38.19	39000	31.13.87.50	3478	Facebook	10617	UDP	64.3	34	2306	33	2736
10/07/19	00:01:08	8869XXXXXXXX	100.100.38.19	39000	157.240.3.52	3478	Facebook	10617	UDP	64.3	3	510	2	176
10/07/19	00:01:08	8869XXXXXXXX	100.100.38.19	39000	31.13.82.48	3478	Facebook	10617	UDP	64.3	3	510	2	510
10/07/19	00:01:08	8869XXXXXXXX	100.100.38.19	59872	157.240.24.52	3478	Facebook	10617	UDP	64.3	3	510	2	176
10/07/19	00:01:08	8869XXXXXXXX	100.100.38.19	59872	157.240.25.52	3478	Facebook	10617	UDP	64.3	3	510	2	176
10/07/19	00:01:18	8869XXXXXXXX	100.100.38.19	59872	1.200.194.62	39668	VIBO	10617	UDP	55.4	725	119625	718	109759
		... .												
10/07/19	00:05:52	8869XXXXXXXX	100.100.38.19	40231	211.77.51.17	443	Facebook	10617	TCP	120.1	23	2894	34	24538
10/07/19	00:05:52	8869XXXXXXXX	100.100.38.19	42392	31.13.87.1	443	Facebook	10617	TCP	94.4	100	15257	81	40312
10/07/19	00:05:52	8869XXXXXXXX	100.100.38.19	46004	31.13.87.15	443	Facebook	10617	TCP	94.5	1315	100583	3707	4611539
10/07/19	00:05:53	8869XXXXXXXX	100.100.38.19	42394	31.13.87.1	443	Facebook	10617	TCP	139.2	40	21842	34	5074

FIGURE 4. Example of ICRs and proposed extended fields [53].



Date	Time	MSISDN	Source IP	Destination IP	Query Domain	Response Address	Response CNAME (If available)
10/07/19	00:05:52	8869XXXXXXXX	100.100.38.19	210.241.208.1	scontent.ftpe1-fna.fbcdn.net	211.77.51.17	—
10/07/19	00:05:52	8869XXXXXXXX	100.100.38.19	210.241.208.1	api.facebook.com	31.13.87.1	star.c10r.facebook.com
* 10/07/19	00:05:52	8869XXXXXXXX	100.100.38.19	210.241.208.1	video.xx.fbcdn.net	31.13.87.15	—
10/07/19	00:05:53	8869XXXXXXXX	100.100.38.19	210.241.208.1	graph.facebook.com	10617	api.facebook.com /star.c10r.facebook.com

FIGURE 5. Example of ICRs with regard to DNS queries and responses.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3009010, IEEE Access

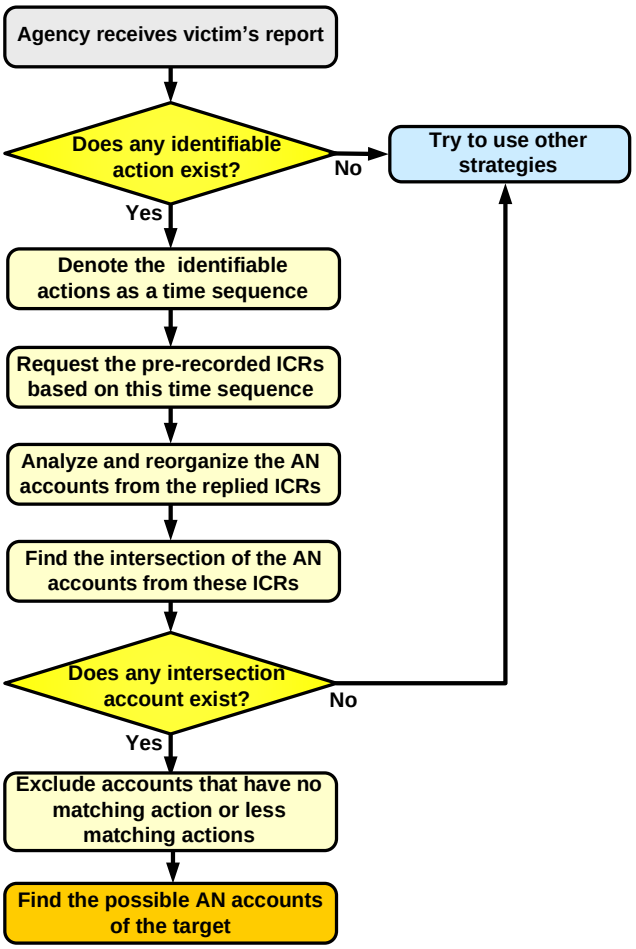


FIGURE 6. Workflow of post-actions sequence mapping strategy.

Although it is possible that several users use the same MSS simultaneously, information from reporters or victims can be used to distinguish actions of the target.

Step 4: Analyze and reorganize requested ICRs, as demonstrated in Fig. 4 and 5, and combine the information with the original sequence.

Step 5: Determine the intersection of the access network (AN) accounts (such as a cell phone number or circuit identification) and the IP addresses from collected ICRs, as shown in Fig. 8(a).

Step 6: After excluding all accounts whose actions do not match, the remaining accounts are the possible targets. Fig. 8(b) shows an example of this.

2) REAL-TIME MEASUREMENT

Almost all the MSSs support nomadic operations. This means that users can use these services anywhere, provided they can access the Internet. When using the previous strategy, it can be occasionally difficult to identify users for nomadic operations because the actions occurred across multiple ANs. For instance, a suspect used WeChat to send a terrorist attack notification from multiple different locations via free WiFi access points. This makes it difficult to correlate the single user identification because there are numerous users with the same action-characteristics. This problem can be overcome using the proposed real-time measurement strategy.

The workflow of this strategy is illustrated in Fig. 9, and its steps are as follows:

Step 1: When a government agency receives information from a victim's report, find any MSS account

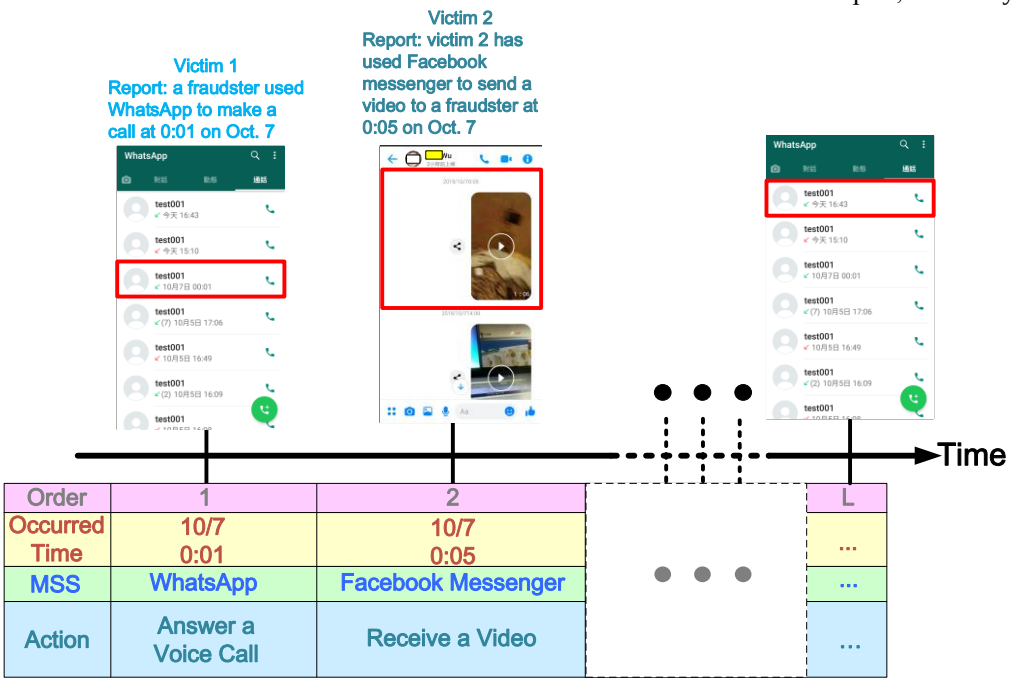


FIGURE 7. Example of the victims' reports received: the suspect used the MSSs during the period.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3009010, IEEE Access

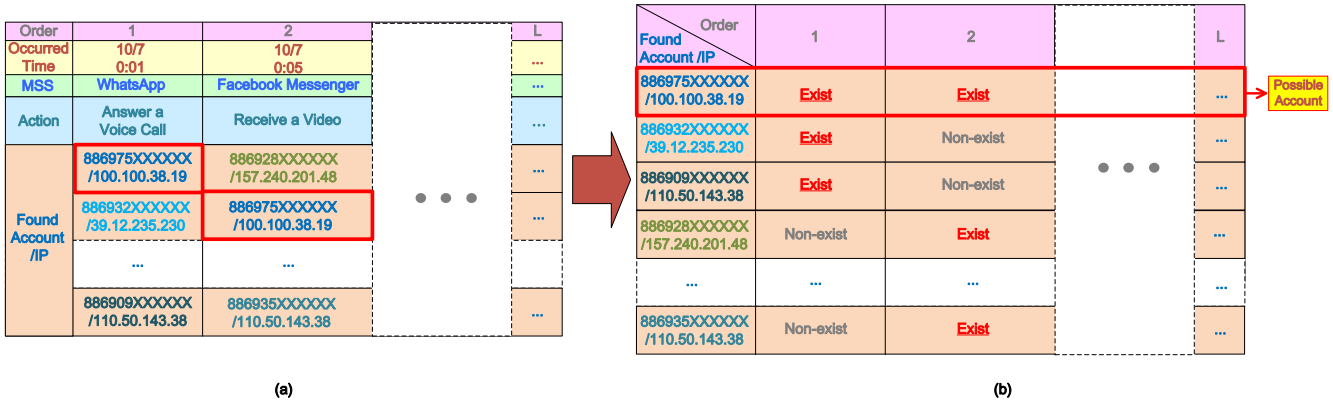


FIGURE 8. Example of applying ICRs: (a) analyze and reorganize ICRs; (b) determine the intersection of accounts using the ICRs.

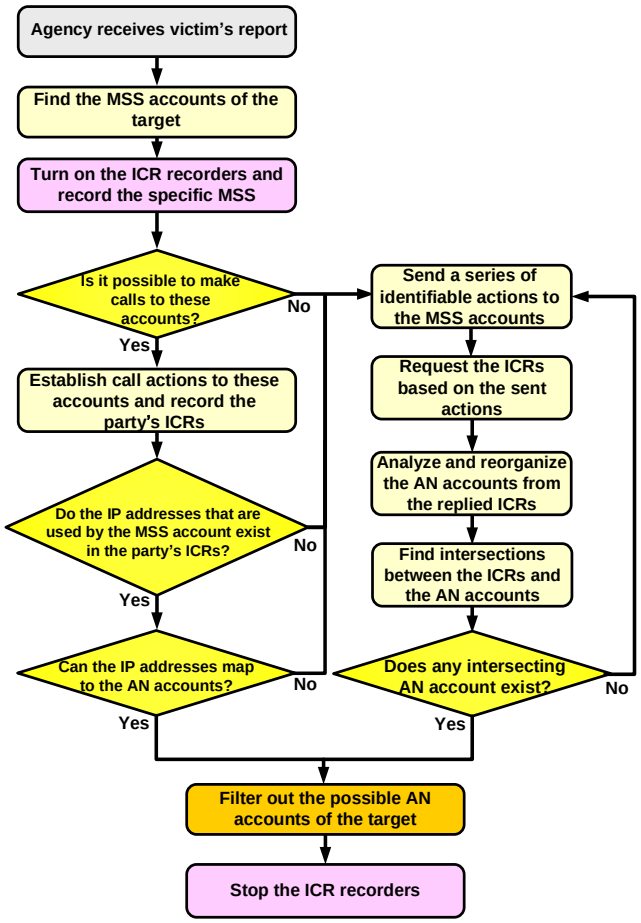


FIGURE 9. Workflow of real-time measurement strategy.

linked to the suspect in the report.

Step 2: Initialize the ICR recorders and record the specific MSS.

Step 3: Check that the voice/video call actions can be used in discovered the target's MSS accounts.

Step 4-1: The call actions can typically be used to disclose the IP addresses of the caller and the callee, provided the callee answers the call. If the discovered accounts have a call action, attempt to establish the call using these accounts.

Step 4-2: Attempt to determine the IP addresses of the call party in the collected ICRs from one side during the call.

Step 4-3: If the IP addresses of the call parties are available, the discovered IP addresses may be verified and mapped to the accounts of the AN by the telecommunication carriers and the IASPs.

Step 5-1: If the call action cannot be reached, the IP addresses of the call parties cannot be found, or the found IP addresses cannot be mapped to the AN accounts, attempt to send a few different messages or call requests to the MSS account during a (short) period of time.

Step 5-2: The same characteristic order can be found in the ICRs of the target from our previous result. The actions sent by government agencies can also be presented as a specific sequence, which is similar to that in Fig. 7. Government agencies can try to request ICRs based on the sent sequence.

Step 5-3: Analyze and reorganize the replied ICRs and combine the information with the original sequence, as shown in Fig. 8(a).

Step 5-4: Determine the intersection of AN accounts from these ICRs, similar to the procedure in Fig. 8(b).

Step 6: Filter out possible AN accounts using the best-matched sequence. These accounts may be potential targets.

Step 7: Stop ICR recorders.

This strategy actively contacts the target. It tries to establish a direct connection between the government agency and the target or make the ICR of the target generate a sequence with specific identifiable characteristics within a short period. As the ICR recorders only function during a short period of time, this strategy can reduce the impact on privacy. In addition, obtaining the current IP address of the target is possible, and it also reduces the effects of nomadic access used between different networks.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3009010, IEEE Access

3) PRACTICAL APPLIED FIELDS

There are two strategies that can be applied to the ICR in order to correlate the MSS account and identify the possible account of a user's AN; this can further be used when corresponding with his/her location. However, in practice, the strategy selected by a government agency depends on practical situations. For instance, in criminal investigations, the LEA aims to avoid alerting the suspects initially as this will hinder their investigation. In this situation, the post-actions sequence mapping strategy is preferable. On the contrary, for suicide or rescue cases, the PSD needs to find or contact the person immediately; thus, the real-time measurement strategy is the preferred solution.

C. EVALUATION OF IDENTIFICATION PERFORMANCE

It is assumed that there are N types of characteristic actions that can be identified and recorded via the ICR recorders. In addition, the probability of each action occurring on the Internet is equal. If a government agency has received records of used MSS (e.g., a chat record from a victim or actively generated identifiable actions) and they determine several identifiable actions at L different times, there are $(N+1)^L$ combinations in this sequence. The probability of each combination can be calculated as $1/(N+1)^L$. A lower probability implies that it is more difficult to find the same matched sequence. In addition, government agencies can increase the MSS action identification capacity or find more identifiable actions at different times in order to reduce the probability of each combination. The relationship between the number of identifiable actions found and the probability of each combination are presented in Fig. 10. In real-world applications, this is useful for excluding large numbers of non-target accounts.

A perfectly matched sequence between the victim's report and the received ICRs is the best-case scenario. However, due to packet loss or other such reasons, ICR recorders may not record the action of the user correctly. This makes it difficult to perfectly match the sequence from the victim's report with the characteristics of the ICRs. Hence, the error tolerance capacity needs to be discussed further.

The error probability of ICR recorders is assumed to be identical, and it is denoted as p. The procedure in Fig. 8(b) is used to set up the error probability model of the ICR recorder, as shown in Fig. 11; x_n, where $n = 0, 1, 2, \ldots, N$, is denoted as identifiable actions of the ICR recorder.

Considering a common decision rule, the majority rule, and assuming that the government agency found k actions that matched between the victim's report sequence and replied ICRs and $k > L/2$, the error probability of the target can be calculated as demonstrated in (2).

$$P_e = P[\,more\ than\ k\ actions\ recorded\ incorrectly\,] = \sum_{i=k+1}^{L} \binom{L}{i} p^i (1-p)^{L-i} \quad (2)$$

The probability of incorrectly identifying a target based on different error probabilities of the ICR recorder and the number of identifiable actions at different times are shown in Fig. 12.

As the majority rule is used, slightly more than $L/2$

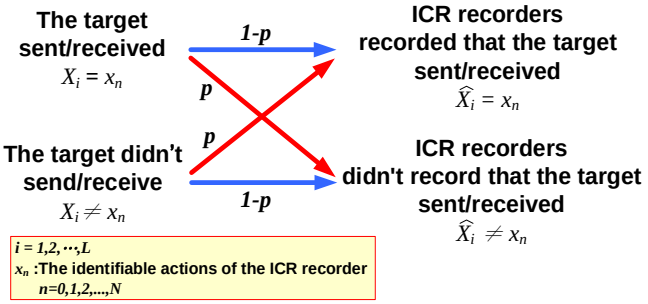


FIGURE 11. Error probability model of the ICR recorder.

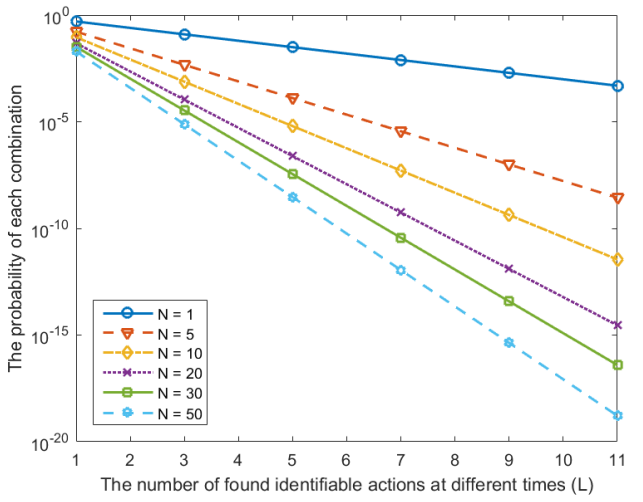


FIGURE 10. Relationship between ICR-identifiable actions (N), number of agencies that found identifiable actions at different times (L), and probability of each combination.

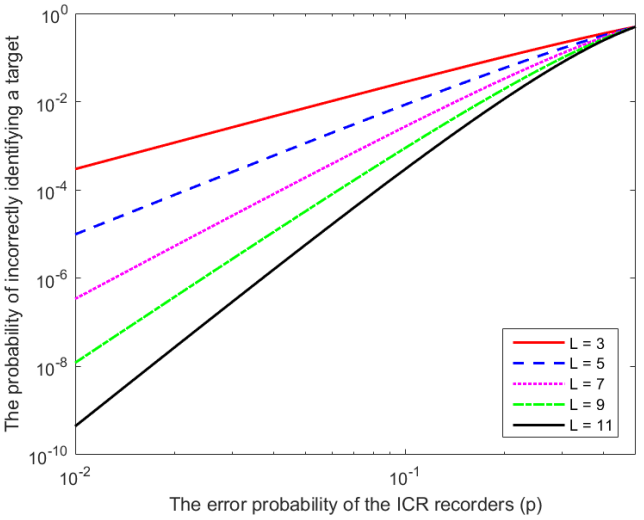


FIGURE 12. Probability of incorrect target identification based on different error probabilities (0.01 ≤ p ≤ 0.5) of ICR recorders and number of identifiable actions found at L different times (using the majority rule).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3009010, IEEE Access

actions were matched, and the users were also denoted as potential targets. A small L can produce a false positive. The users that satisfy the majority rule and match more actions at the same time have a higher probability of being the potential target. Under these conditions, the error probability can be reduced. Fig. 13 depicts the relationship between error probability and the mismatched number of actions.

D. COMPARISON OF DIFFERENT ICR RETENTION FRAMEWORKS

This section presents a comparison of some existing ICR retention frameworks and the proposed framework. The compared frameworks include the well-known Danish [54]–[56] and United Kingdom[15], [32], [53], [56] data retention frameworks as well as special frameworks that use the MITM mechanism [41]; this is because they provide more transparency in the information. The comparison of these frameworks is presented in Table XII.

Based on the goals, retained range and data, real application issues, and identification performance of these frameworks, the proposed framework offers some excellent properties; it is also more suitable when applied to the MSSs.

When considering the characteristics of MSSs, only the information of the session related to the MSSs should be retained; this down sizes the volume and range of retained traffic and also reduces the number of potential users. It is also available without pre-storing ICRs (with the real-time measurement strategy). Our framework makes it possible to correlate both communication parties. The evaluation of identification performance, which has not been discussed thoroughly for the other frameworks, is achievable using our framework. Apart from the abovementioned advantages, our framework is also applicable to some broadly discussed issues [56], such as security of fully retained data and privacy risks, encryption traffic, constant connections, and complex processing to create and store ICRs.

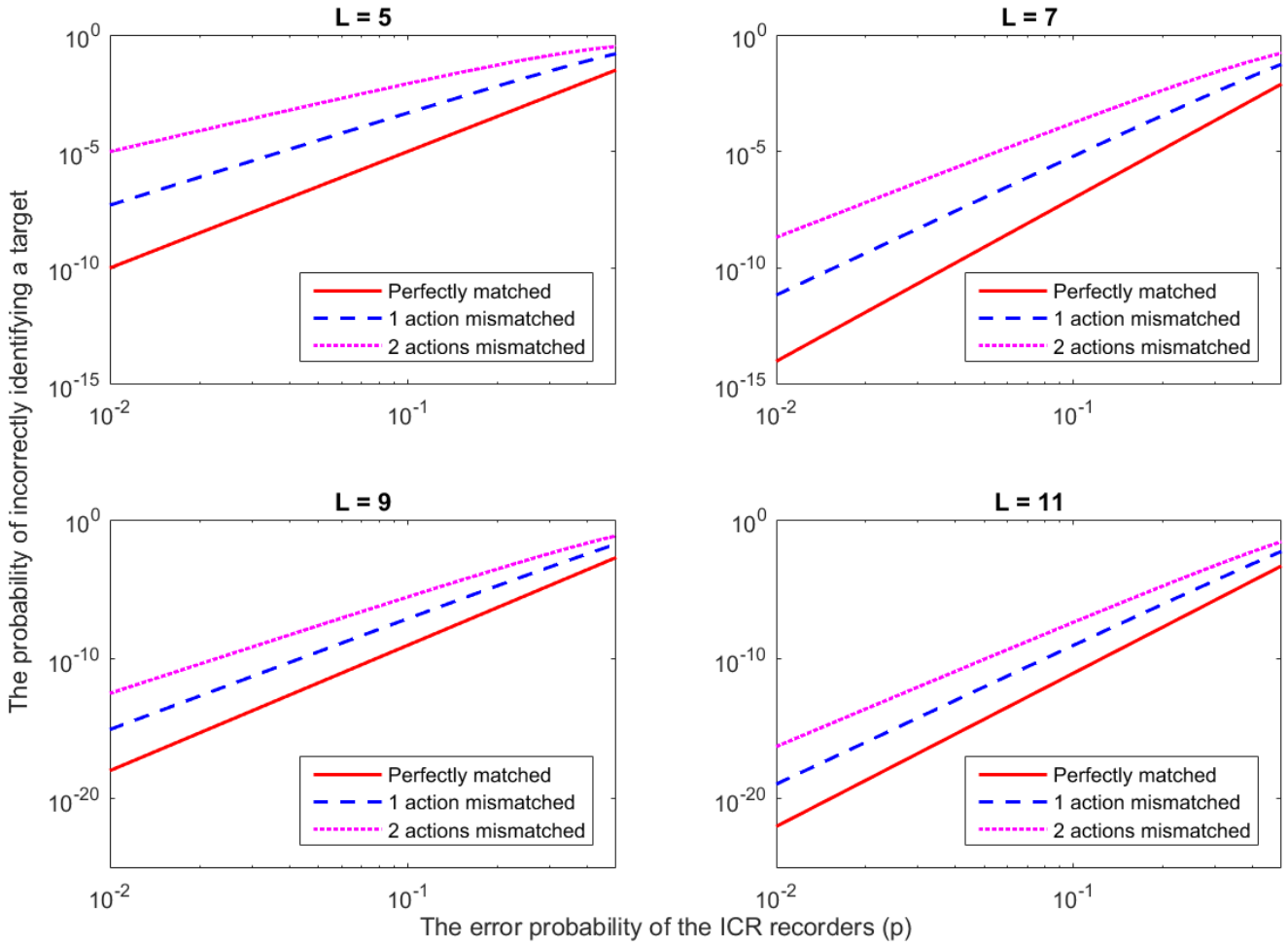


FIGURE 13. Probability of identification error with the number of matched actions.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3009010, IEEE Access

TABLE XII
COMPARISON OF DIFFERENT RETENTION FRAMEWORKS FOR INTERNET CONNECTION RECORDS

Framework (countries or authors and year)	Danish data retention (Denmark, 2007-2014, 2016)	Internet connection records in the Investigatory Powers Act 2016 (The United Kingdom, 2016)	Cloud-based forensics tracking scheme for online social network clients (Lin et al., 2015)	Proposed framework (Li et al., 2020)
Goal	Retaining and storing traffic data is possible when using these data in conjunction with investigation and prosecution of criminal offences.	1. To identify the sender. 2. To identify the communication services used. 3. To determine whether illegal material has been accessed or created online.	To determine the actual identity of criminal suspects and their geolocation through online social networks (OSNs).	1. To identify the sender of an MSS communication. 2. To identify the MSS used. 3. To determine whether illegal material has been accessed or created online on a specific MSS. 4. To provide fast responses to MSS users during emergencies.
Retention range and data	1. Two retention approaches: (1) First and last packet of each session. (2) Every 500th packet for the boundaries of the ISP. 2. Source/destination IP addresses and port number, transmission protocol, and timestamps of the session are retained. 3. The internal traffic within the ISP's own network, such as the DNS lookups, are excluded for the retention range. 4. Deep packet inspection (DPI) equipment is not applied.	1. A clear retention range is not indicated (probability of the entire traffic for all users). 2. Date, time, Internet service identification of the user (e.g. mobile phone number), source/destination IP addresses and port number, service (or domain), and location of the session should be retained.	1. Only information of the session related to the OSNs should be retained. 2. User information with timestamps should be retained. (1) Internet access. (2) Private/public IP mapping. (3) Source/destination IP addresses and port numbers. (4) OSNs' login information (such as the OSN accounts with the IP addresses and port numbers used) obtained via man-in-the-middle (MITM).	1. Only information of the session related to the MSSs should be retained. 2. This framework includes the UK proposed retained data and also incorporates the transmission protocol, volume of outgoing/incoming traffic, and the session. 3. The IP address and Internet service identification of the user requesting specific domain names of MSSs and the responses of domain name servers (i.e., the IP addresses of MSS servers) are recorded with a timestamp.
Issues hindering practical application for MSSs field	1. Excessive information is retained (the Danish government indicated that 3,500 billion telecommunication records were retained in 2013). 2. Several packets with interactive traffic, such as instant messaging or online gaming, are smaller. If only the 500th packet at the boundaries of the ISP are retained, significant amounts of useful information may be lost. 3. It does not evaluate the relationship between MSS communication parties. 4. Considers multiple services host on the same IP address. It is not possible to know which services were accessed without any additional information. 5. The Danish government has decided to repeal this framework because it was unable to achieve the stated objective (investigation and prosecution of crime).	1. All sessions for all Internet users are recorded as ICRs. The retained ICRs are large and most of them are useless. 2. When applied to the MSS field, huge sessions are produced (there are more sessions when the device is idle), making it difficult to identify the intersection between users. 3. It does not evaluate the relationship between MSS communication parties. 4. If the retained ICRs include a service name or a web address (e.g. www.facebook.com), the DPI is required in this framework.	1. Retained information is duplicated, massive, and complex. 2. A pre-plugged custom or self-signed certificate into the users' certificate store is necessary for the MITM mechanism, as it is a workable purpose. It induces critical impacts on security and privacy. 3. If applied to popular MSSs, huge sessions are produced, making it difficult to identify the intersection between users. 4. It does not evaluate the relationship between MSS communication parties. 5. High deployment costs due to high-speed MITM and DPI servers.	1. Based on known characteristics of MSS actions, all retained information is only related to the MSSs. This reduces the volume and range of retained information. 2. Developed strategies, i.e., post-action sequence mapping and real-time measurement strategies, can reduce the number of possible users and increase the probability of accurate identification. The real-time measurement strategy is applied without pre-retention of the ICRs by the IASPs. 4. It can be applied to encrypted MSSs without breaking down the original security mechanism. 5. It can correlate the communication parties of the MSS. 6. The IP addresses of MSS servers are often known in advance because they are key characteristics for the retention in the framework. Hence, this framework functions without requiring DPI. 7. If a new MSS is introduced or the characteristics of the original MSS are altered, the new characteristics must be identified prior to retention.
Identification performance	No further discussion.	No further discussion.	No further discussion.	It can be evaluated.

V. CHALLENGES AND FUTURE RESEARCH

This study proved that ICRs offer the potential of new opportunities for governments, especially in emergency events, criminal investigations, and anti-terrorist agencies; it also presents novel security and privacy impacts. However, further research is required to address some challenges in real-world applications [32], [57]. This section highlights a few important technological and management challenges and potential scope of future research.

A. IDENTIFICATION EFFICIENCY

Google Play and Apple App, the largest mobile application stores, provided more than 2.2 million applications each in 2017 [58]. These applications facilitate instant multimedia messaging and voice communication functions, while being subjected to rapid updates and changes. In addition, different devices and operating systems, such as Google's Android, Apple's iOS, Blackberry, and Windows mobile, are currently used in the global smartphone market [59].

Considering this abundance of mobile applications and devices, governments need additional resources to research the action traffic for each application. Therefore, it is vital to develop methods to apply artificial intelligence (AI) and machine learning to overcome this problem.

B. TRAFFIC ENCRYPTION AND SECURE CONNECTION

For security and privacy issues, most MSS services are not only equipped with an encryption mechanism, but they can also use secure connections to access these services, such as Tor [60] or virtual private networks (VPNs). As a result, ICR recorders may not be able to recognize traffic behaviors easily [56]. It is considerably challenging to overcome such issues in practical applications.

C. PRIVACY IMPACT

For emergency services and public security, governments typically engage in massive information gathering and processing. Citizens are typically concerned about the transparency, collection, processing, retention, and distribution of their data. However, details of retained data are not disclosed to the public. If such data are disclosed, offenders and terrorists can develop methods to evade government agencies. Hence, governments should research and develop control, management, and secure retention mechanisms, while considering a suitable transparency framework to achieve a balance between public security and privacy. A few specific issues need to be discussed, such as the necessity of a national-level transparency institution, overseeing jobs from technological and legal perspectives, verifying the legality of government-retained proposals, managing oversight and accountability, determining if retained data is insufficient or excessive, ensuring government agencies access necessary ICRs in an authorized and limited manner, securing retained ICRs to avoid unauthorized access, secure delivery of ICRs to PSDs and LEAs.

D. COST

For ICR retention, ICR recorders must be deployed all the networks of Internet service providers in a country; however, this is significantly expensive. For instance, in a long-term evolution network (LTE), recorders are located between the eNodeB and the serving gateway (SGW), between the SGW and the packet data network gateway (PDN GW, PGW), or between the PGW and PDN. In addition, there are different deployment and maintenance costs. Thus, designing the architecture of an ICR recorder and planning an optimal overall retention network to reduce deployment costs is one of the most important goals in this field.

VI. CONCLUSIONS

This study investigated the potential of ICRs in tracking MSS users in law enforcement and rescue fields through an ICR retention framework; by recording, analyzing, and comparing characteristic ICRs of MSSs, this study demonstrated that general and indiscriminate pre-retention of ICRs is potentially no longer necessary in the abovementioned fields. This can be substituted by the actions-based pre-retention of ICRs.

In addition, we suggest that retained data are mainly based on existing network flow monitoring technologies; thus, these data can be used to correlate the relationship between communication parties and identify possible targets from numerous MSS users. As a significant amount of ICRs with the same activity characteristics are generated simultaneously, it is difficult to precisely identify the target. Accordingly, two identification strategies were introduced in the proposed framework to increase the probability of accurate identification. For one of these proposed strategies, telecommunication carriers and IASPs do not require the pre-retention of ICRs. This strategy can be directly applied to the daily operations of government agencies. This study provides a useful reference for agencies worldwide, including governmental, non-governmental, and civil society organizations.

The abundance of MSS applications and devices, along with the identification efficiency, privacy, and their associated costs, are the primary challenges in employing ICRs; these issues warrant further research to ensure appropriate real-world applications. These research areas are expected to play crucial roles in global data retention legislation and the development of effective management techniques in the future.

REFERENCES

[1] X. Hu, T. H. S. Chu, V. C. M. Leung, E. C.-H. Ngai, P. Kruchten, and H. C. B. Chan, "A survey on mobile social networks: Applications, platforms, system architectures, and future research directions," *IEEE Commun. Surv. Tutorials*, vol. 17, no. 3, pp. 1557–1581, 2015.

[2] Zacks, "Why is Facebook (FB) up 17% since its last earnings report?" May. 2018 [Online]. Available: https://www.nasdaq.com/article/why-is-facebook-fb-up-17-since-its-last-earnings-report-cm968844, Accessed on: Jul. 14, 2018.

[3] "2018 world population by country," 2018. [Online]. Available: http://worldpopulationreview.com/, Accessed on: Jul. 15, 2018.

[4] S. Kemp, "Digital in 2018: World's internet users pass the 4 billion mark," We Are Social, Jan. 30, 2018. [Online]. Available: https://wearesocial.com/blog/2018/01/global-digital-report-2018, Accessed on: Jul. 24, 2018

[5] "Investigatory powers bill factsheet – Internet connection records," *UK Home Office*, July. 8, 2016. [Online]. Available: https://www.gov.uk/government/publications/investigatory-powers-bill-fact-sheets, Accessed on: May 27, 2020.

[6] "Information for law enforcement authorities," *Facebook*. [Online]. Available: https://www.facebook.com/safety/groups/law/guidelines/, Accessed on: Aug. 2, 2018.

[7] "Guidelines for law enforcement," *Twitter*. [Online]. Available: https://help.twitter.com/en/rules-and-policies/twitter-law-enforcement-support, Accessed on: Aug. 2, 2018.

[8] "Tumblr law enforcement guidelines," *Tumblr*. [Online]. Available: https://www.tumblr.com/docs/en/law_enforcement, Accessed on: Aug. 2, 2018.

[9] "WhatsApp FAQ - Information for law enforcement authorities," *WhatsApp*. [Online]. Available: https://faq.whatsapp.com/en/26000050/?category=5245250, Accessed on: Aug. 2, 2018.

[10] "User info disclosure requests by law agencies," *LINE*. [Online]. Available: https://linecorp.com/en/security/article/35, Accessed on: Aug. 2, 2018.

[11] S. Karantzoulidis, "The best and worst police response times of 10 major U.S. cities," *Security Sales & Integration* https://www.securitysales.com/news/best-worst-police-response-times/, Accessed on: Dec. 3, 2019.

[12] C. Hymas and A. Kirk, "Crime victims wait half an hour for police to respond to 999 calls as response times double," Telegraph, Jan. 13, 2019 [Online]. Available: https://www.telegraph.co.uk/news/2019/01/13/crime-victims-wait-half-hour-police-respond-999-calls-response/, Accessed on: Dec. 3, 2019

[13] Telecommunications (Interception and Access) Amendment (Data Retention) Act, Laws of Australian, 2015. [Online]. Available: https://www.legislation.gov.au/Details/C2015A00039, Accessed on: Jul. 24, 2018.

[14] Investigatory Powers Act, Laws of UK, 2016. [Online]. Available: http://www.legislation.gov.uk/ukpga/2016/25/section/62, Accessed on: Jul. 24, 2018.

[15] "Judgment in investigatory powers legal challenge," *The UK Home Office in media*, July 20, 2019. [Online]. Available: https://homeofficemedia.blog.gov.uk/2019/07/29/judgment-in-investigatory-powers-legal-challenge/, Accessed on: May 5, 2020.

[16] N. Suzoru, K. Pappalardo, and N. McIntosh, "The passage of Australia's data retention regime: national security, human rights, and media scrutiny," *Internet Policy Rev.*, vol. 6, no. 1, 2017. Accessed on: Jul. 24, 2018, DOI: 10.14763/2017.1.454.

[17] C. Barker and J. Murphy, "Telecommunications data retention – Budget Review 2015–16 Index," *Parliament of Australia*. [Online]. Available: https://www.aph.gov.au/About_Parliament/Parliamentary_Departments/Parliamentary_Library/pubs/rp/BudgetReview201516/Telco, Accessed on: Aug. 3, 2018.

[18] A. Travis and O. Bowcott, "Telecoms companies raise questions over cost and feasibility of 'snooper's charter," *The Guardian*, Dec. 15, 2015. [Online]. Available: https://www.theguardian.com/world/2015/dec/15/bt-vodafone-o2-ee-3-cost-feasibility-snoopers-charter, Accessed on: Aug. 3, 2018.

[19] D. Campbell, "Britain to pay billions for monster internet surveillance network," *Computer Weekly*, Mar. 21, 2016. [Online]. Available: https://www.computerweekly.com/news/4500279596/Britain-to-pay-billions-for-monster-internet-surveillance-network, Accessed on: Aug. 3, 2018.

[20] "Internet connection record - Digital marketplace." [Online]. Available: https://www.digitalmarketplace.service.gov.uk/digital-outcomes-and-specialists/opportunities/9955, Accessed on: Nov. 1, 2019.

[21] "Internet connection record 2 - Digital marketplace." [Online]. Available: https://www.digitalmarketplace.service.gov.uk/digital-outcomes-and-specialists/opportunities/9956, Accessed on: Nov. 1, 2019.

[22] D. Roberts and S. Ackerman, "Anger swells after NSA phone records court order revelations," *The Guardian*, Jun. 7, 2013 [Online]. Available: https://www.theguardian.com/world/2013/jun/06/obama-administration-nsa-verizon-records, Accessed on: Aug. 6, 2018.

[23] G. Greenwald and E, MacAskill, "NSA Prism program taps in to user data of Apple, Google and others," *The Guardian*, Jun. 7, 2013 [Online]. Available: https://www.theguardian.com/world/2013/jun/06/us-tech-giants-nsa-data, Accessed on: Aug. 4, 2018.

[24] S. Landau, "Making sense from Snowden: What's significant in the NSA surveillance revelations," *IEEE Secur. Priv.*, vol. 11, no. 4, pp. 54–63, Jul. 2013. Accessed on: Aug. 3, 2018 DOI: 10.1109/MSP.2013.90.

[25] S. Landau, "Highlights from making sense of Snowden, Part II: What's significant in the NSA revelations," *IEEE Secur. Priv.*, vol. 12, no. 1, pp. 62–64, Jan. 2014, Accessed on: Aug. 4, 2018, DOI: 10.1109/MSP.2013.161.

[26] S. Siddiqui, "Congress passes NSA surveillance reform in vindication for Snowden," *The Guardian*, Jun. 3, 2015 [Online]. Available: https://www.theguardian.com/us-news/2015/jun/02/congress-surveillance-reform-edward-snowden, Accessed on: Aug. 04, 2018.

[27] Joined Cases C-293/12 and C-594/12 Digital Rights Ireland Ltd (C-293/12) v Minister for Communications Marine and Natural Resources and Others and Kärntner Landesregierung (C-594/12), April 2014, [online] Available: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:62012CJ0293&rid=1, Accessed on: Aug. 5, 2018.

[28] T. Papademetriou, "European Union: ECJ invalidates data retention directive," *The Library of Congress*, 2014. [Online]. Available: http://www.loc.gov/law/help/eu-data-retention-directive/eu.php#_ftn1, Accessed on: Aug. 5, 2018.

[29] Joined Cases Tele2 Sverige AB v Post- och telestyrelsen (C-203/15) and Secretary of State for the Home Department v. Watson (C-698/15), Dec. 2016, [Online]. Available: http://curia.europa.eu/juris/document/document.jsf?text=&docid=186492&pageIndex=0&doclang=en&mode=lst&dir=&occ=first&part=1&cid=17504, Accessed on: Aug. 5, 2018.

[30] O. Bowcott, "EU's highest court delivers blow to UK snooper's charter," *The Guardian*, Dec. 21, 2016. [Online]. Available: https://www.theguardian.com/law/2016/dec/21/eus-highest-court-delivers-blow-to-uk-snoopers-charter, Accessed on: Aug. 6, 2018.

[31] S. Carey, "This is how the police want your internet connection records to look," *Computer World UK*, Jun. 8, 2016. [Online]. Available: https://www.computerworlduk.com/security/this-is-how-police-want-your-internet-connection-records-look-3641581/, Accessed on: Aug. 9, 2018.

[32] "Draft investigatory powers bill - Joint committee on the draft investigatory powers bill," *UK Parliament*, 2015. [Online]. Available: https://publications.parliament.uk/pa/jt201516/jtselect/jtinvpowers/93/9307.htm, Accessed on: Aug. 16, 2018.

[33] "Vodafone group plc annual report 2017," *Vodafone*, London, England, UK, Rep., Mar. 2017.

[34] C. Smith, "25 amazing AT&T statistics and facts (June 2018)," *DMR*, Jun. 27, 2018. [Online]. Available:

https://expandedramblings.com/index.php/att-statistics/, Accessed on: Aug. 10, 2018.

[35] L. Onwuzurike and E. DeCristofaro, "Experimental analysis of popular anonymous, ephemeral, and end-to-end encrypted apps," presented at *NDSS Symposium 2016*, San Diego, California, USA, Feb. 21-24, 2016.

[36] M. Conti, L. V. Mancini, R. Spolaor, and N. V. Verde, "Analyzing android encrypted network traffic to identify user actions," *IEEE Trans. Inf. Forensics Secur.*, vol. 11, no. 1, pp. 114 – 125, Jan. 2016, DOI: 10.1109/TIFS.2015.2478741.

[37] M. Conti, Q. Q. Li, A. Maragno, and R. Spolaor, "The dark side(-channel) of mobile devices: A survey on network traffic analysis," *IEEE Commun. Surv. Tutorials*, vol. 20, no. 4, pp. 2658 – 2713, 2018, DOI: 10.1109/COMST.2018.2843533.

[38] M. Swarnkar, N. Hubballi, N. Tripathi and M. Conti, "AppHunter: Mobile application traffic classification," *2018 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*, Indore, India, 2018, pp. 1-6, DOI: 10.1109/ANTS.2018.8710170.

[39] V. F. Taylor, R. Spolaor, M. Conti, and I. Martinovic, "Robust smartphone app identification via encrypted network traffic analysis," *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 1, pp. 63 – 78, Jan.2018, DOI: 10.1109/TIFS.2017.2737970.

[40] D. Li, W. Li, X. Wang, C. T. Nguyen, and S. Lu, "ActiveTracker: Uncovering the trajectory of app activities over encrypted internet traffic streams," in Annual *IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks workshops*, Jun. 2019, DOI: 10.1109/SAHCN.2019.8824928.

[41] F.-Y. Lin, C.-C. Huang, and P.-Y. Chang, "A cloud-based forensics tracking scheme for online social network clients," *Forensic Sci. Int.*, vol. 255, pp. 64–71, Oct. 2015. DOI: 10.1016/j.forsciint.2015.08.011.

[42] C. Wass, "Four ways to bypass android SSL verification and certificate pinning," *NetSPI Blog*, Jan. 9, 2018. [Online]. Available: https://blog.netspi.com/four-ways-bypass-android-ssl-verification-certificate-pinning/, Accessed on: Aug. 14, 2018.

[43] R. Valsky, "Certificate pinning on mobile applications," *Perimeterx*, Mar. 6, 2018. [Online]. Available: https://www.perimeterx.com/blog/certificate-pinning-on-mobile/#, Accessed on: Aug. 14, 2018.

[44] Android tcpdump [Online]. Available: https://www.androidtcpdump.com/, Accessed on: Jan. 10, 2020.

[45] "Recording a packet trace," *Apple Developer Documentation*, [Online]. Available: https://developer.apple.com/documentation/network/recording_a_packet_trace, Accessed on: Feb. 28, 2020.

[46] Arne Holst, "Mobile OS market share 2019," *Statista*, Sep. 13, 2019. [Online]. Available: https://www.statista.com/statistics/272698/global-market-share-held-by-mobile-operating-systems-since-2009/, Accessed on: Dec. 30, 2019.

[47] B. Bucher, "Messaging app usage statistics around the world," MessengerPeople, 2019. [Online]. Available: https://www.messengerpeople.com/global-messenger-usage-statistics/, Accessed on: Feb. 22, 2019.

[48] M. W.-Wilson. "WhatsApp vs telegram vs signal." TechRadar. https://www.techradar.com/news/whatsapp-vs-telegram-vs-signal, Accessed on: Jan. 11, 2020.

[49] A. Yorgan, "Social media usage in Russia: 10 key statistics | Russian Search Marketing," Russian Search Marketing, Mar. 19, 2019. [Online]. Available: https://russiansearchmarketing.com/10-key-statistics-social-media-usage-russia-2019/, Accessed on: Jan. 10, 2020.

[50] "iMessage and FaceTime & Privacy," *Apple Support*, [Online]. Available: https://support.apple.com/en-us/HT209110, Accessed on: Mar. 03, 2020.

[51] A. Callado *et al.*, "A survey on internet traffic identification," *IEEE Commun. Surv. Tutorials*, vol. 11, no. 3, pp. 37–52, 2009. Accessed on: Nov. 13, 2018, DOI: 10.1109/SURV.2009.090304.

[52] R. Hofstede *et al.*, "Flow monitoring explained: From packet capture to data analysis with NetFlow and IPFIX," *IEEE Commun. Surv. Tutorials*, vol. 16, no. 4, pp. 2037–2064, 2014. Accessed on: Nov. 13, 2018, DOI: 10.1109/COMST.2014.2321898.

[53] "Joint law enforcement written submission to the joint committee on the draft investigatory powers bill," National Police Chiefs' Council, Feb. 2016. [Online]. Available: https://www.npcc.police.uk/Publication/IPB JC LE Written Submission Annexes A to D v1.pdf, Accessed on: Sep. 18, 2019.

[54] "Danish administrative order for data retention (Logningsbekendtgørelsen)," *Denmark Ministry of Justice*, 2006. [Online]. Available: https://www.retsinformation.dk/eli/lta/2006/988, Accessed on: May. 16, 2020.

[55] "Published written evidence (IPB0051) of Investigatory Powers Bill," *UK Parliament*, 2015. [Online]. Available: http://data.parliament.uk/WrittenEvidence/CommitteeEvidence.svc/EvidenceDocument/Science and Technology/Investigatory Powers Bill Technology issues/written/25190.html, Accessed on: May 8, 2020.

[56] "Comparison of internet connection records in the investigatory powers bill with Danish internet session logging legislation," *GOV.UK*, Feb. 2016. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/504189/Comparison_of_ICRs_with_Danish_Session_Logging.pdf Accessed on: May 16, 2020.

[57] S. Trendall, "Government to trial scheme to record citizens' internet connections," *Public Technology.net*, Jun. 26, 2019. [Online]. Available: https://www.publictechnology.net/articles/news/government-trial-scheme-record-citizens'-internet-connections, Accessed on: May 28, 2020.

[58] "App stores: number of apps in leading app stores 2017," *Statista*, Mar. 2017. [Online]. Available: https://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores/, Accessed on: Dec. 30, 2017.

[59] "Smartphone growth expected to remain positive as shipments forecast to grow to 1.7 billion in 2021, according to IDC," *IDC*, Aug. 29, 2017 [Online]. Available: https://www.idc.com/getdoc.jsp?containerId=prUS43010517, Accessed on: Jan. 23, 2018.

[60] "Tor: Overview," *Tor project*. [Online]. Available: https://2019.www.torproject.org/about/overview.html.e, Accessed on: Nov. 07, 2019.

CHEN-YU LI received his B.S.E.E. degree from Chung Yuan Christian University, Taoyuan, Taiwan, R.O.C., in 2005, and his M.S. degree in Electronic and Computer Engineering from the National Taiwan University of Science and Technology, in 2007. He is currently pursuing his Ph.D. degree with the Graduate Institute of Communication Engineering and the Department of Electrical Engineering at the National Taiwan University. His research interests include optical and wireless communication systems, computer networks, digital forensics, anonymity and privacy, lawful interception, and network security.

CHIEN-CHENG HUANG received his Ph.D. degree in Information Management from the National Taiwan University in 2014. His current research interests include data mining, business intelligence, information security, and cyber/network forensics.

FEIPEI LAI received his B.S.E.E. degree from the National Taiwan University, in 1980, and his M.S. and Ph.D. degrees in Computer Science from the University of Illinois at Urbana–Champaign, in 1984 and 1987, respectively. He is currently a Professor with the Graduate Institute of Biomedical Electronics and Bioinformatics, the Department of Computer Science and Information Engineering, and the Department of Electrical Engineering, National Taiwan University. He was a Vice Superintendent of the National Taiwan University Hospital, the Chairman of the Taiwan Network Information Center, and a Visiting Professor with the Department of Computer Science and Engineering at the University of Minnesota in Minneapolis, MN, USA. He was also a Guest Professor at the University of Dortmund in Germany, and a Visiting Senior Computer System Engineer for the Center for Supercomputing Research and Development at the University of Illinois at Urbana–Champaign. He currently holds ten Taiwan patents and four U.S. patents.

SAN-LIANG LEE (SM'07) received his Ph.D. degree in Electrical and Computer Engineering from the University of California, Santa Barbara (UCSB), in 1995. He joined the faculty of the Department of Electronic Engineering at the National Taiwan University of Science and Technology (NTUST) in 1988 and became a Full Professor in 2002. He served as the Vice President of the university from 2011 to 2014. He was the Chairman of the department from 2005 to 2008. He has also served as the Dean of the Academic Affairs Office, NTUST, from 2008 to 2010. He was the director of the program office for the National Innovative Education Program on Image Display Technology, sponsored by the Ministry of Education, Taiwan, from 2005 to 2009. He served as the Electronic Section Editor of the SCI indexed Journal of the Chinese Institute of Engineers from 2007 to 2012. He was also a visiting scientist in the Research Laboratory of Electronics at the Massachusetts Institute of Technology (MIT), and he took a sabbatical leave from NTUST from 2010 to 2011. His research interests include semiconductor optoelectronic components, photonic integrated circuits, nanophotonics, and optical networking technologies. He has published more than 200 refereed papers in international journals and conferences and holds 30 patents.

JINGSHOWN WU (M'78–SM'99–F'05–LF'09) received his B.S. and M.S. degrees in Electrical Engineering from the National Taiwan University, Taipei, Taiwan, R.O.C., and his Ph.D. degree from Cornell University, Ithaca, NY, in 1970, 1972, and 1978, respectively. He joined Bell Laboratories in 1978, where he worked on digital network standards and performance, and optical fiber communication systems. In 1984, he joined the Department of Electrical Engineering at the National Taiwan University as a Professor and he was the Chairman of the department from 1987 to 1989.

He was also the Director of the Communication Research Center at the College of Engineering of the university from 1992 to 1995. From 1995 to 1998, he served as the Director of the Division of Engineering and Applied Science, National Science Council, R.O.C., while on leave from the university. From 1999 to 2002, he acted as the Chairman of the Commission on Research and Development, and the Director of the Center for Sponsor Programs at the National Taiwan University. He was the Vice-President of the university from 2002 to 2005. He was the Chairman of the Institute for Information Industry from 2006 to 2007. He is interested in optical fiber communications, computer communications, and communication systems. He has published more than 160 journal and conference papers and holds 16 patents. Professor Wu received the Distinguished Research Awards and became a Distinguished Research Fellow from the National Science Council R.O.C from 1991 to 1996 and from 1996 to 2002. He was the recipient of the outstanding engineering professor award from the Chinese Institute of Engineers in 1996. He also received the engineer metal award from the Institute of Chinese Electrical Engineers in 2006, and the award from CIE/USA in 2009, respectively.

Professor Wu is a life fellow of IEEE, a life member of the Chinese Institute of Engineers, the Optical Society of China, and the Institute of Chinese Electrical Engineers. He served as the Vice Chairman (1997–1998) and the Chairman (1998–2000) of the IEEE Taipei Chapter. He was also a member of the IEEE Communications Society Award Committee from 2006 to 2008, the IEEE Communications Society Fellow Evaluation Committee 2008, and the IEEE Fellow Committee from 2009 to 2012.

RONG-CHI CHANG received his M.A. and Ph.D. degrees in Computer Science and Information Engineering from Tamkang University, Taiwan.

He is currently an Associate Professor and the head of the Department of Technology Crime Investigation at the Taiwan Police College in Taiwan. His research interests include digital content analysis, multimedia security, digital image processing, pattern recognition, and digital crime scene analysis.

Hsiang-Wei Huang received his Bachelor's degree in Information Management from the Central Police University in 2011, and a Master's degree in Computer Science from the National Chengchi University in 2017. His research interests include cyber security, malware analysis, cryptocurrency tracking, and digital forensics for computer, mobile, cloud, and the Internet of Things (IoT).