IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 29, 2018, accepted May 24, 2018, date of publication May 29, 2018, date of current version June 26, 2018.

*Digital Object Identifier 10.1109/ACCESS.2018.2841878*

# A Smart Automated Signature Extraction Scheme for Mobile Phone Number in Human-Centered Smart Home Systems

**PAN WANG[1,2], (Member, IEEE), XUEJIAO CHEN[3], (Member, IEEE), FENG YE[4], (Member, IEEE), AND ZHIXIN SUN[1,2]**

[1]School of Modern Posts, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
[2]Institute of Modern Posts, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
[3]School of Communication, Nanjing College of Information Technology, Nanjing 210023, China
[4]Department of Electrical and Computer Engineering, University of Dayton, Dayton, OH 45469, USA

Corresponding author: Pan Wang (wangpan@njupt.edu.cn)

This work was supported in part by the 2016 Jiangsu Provincial Government Scholarship Program, China, in part by the 2017 Jiangsu Overseas Visiting Scholar Program for University Prominent Young & Middle-Aged Teachers and Presidents, China, in part by the Top-Notch Academic Programs Project of Jiangsu Higher Education Institutions under Grant PPZY2015A092, in part by the National Natural Science Foundation of China under Grant 61373135, Grant 61672299, Grant 61702281, and Grant 61602259, and in part by the Natural Science Foundation of Jiangsu Province under Grant BK20150866 and Grant BK20160913.

**ABSTRACT** Human-centered smart devices profiling with Wi-Fi networks has received much attention from both research and industry, especially those network operators and security agencies who aim to enhance user experience and security of home network as well as free Wi-Fi services. One type of such profiling is the extraction of mobile phone numbers. In traditional cellular networks, such as 3G and 4G, mobile phone number extraction can be achieved from the analysis of the authentication signaling. However, this method cannot be used in the broadband network environment, e.g., Wi-Fi. Operators and security agencies of Wi-Fi networks often apply manual statistics, telephone inquiries or user input for information. Unfortunately, those traditional methods are inefficient in practice. Moreover, authenticity cannot be guaranteed with the traditional methods. In this paper, we propose a smart method for mobile phone number extraction in smart home networks and systems. In particular, the proposed method is based on deep packet inspection of home broadband traffic. To improve the efficiency and accuracy of detection, we further propose a smart automated signature extraction method of mobile phone numbers from home network traffic. Our proposed method can achieve 86.2% accuracy in the real-life human-centered smart home network test.

**INDEX TERMS** Automated signature extraction, smart home, deep packet inspection, mobile phone number.

## I. INTRODUCTION

With the development of Smart Home, more and more Human-centered smart devices, especially mobile phones, access the Internet via Wi-Fi [1]. In order to achieve security, control and fine operation of networks, it is important to extract some information such as the mobile phone numbers by analyzing and identifying the network traffic efficiently and accurately [2]. As all to know, mobile phones need to send information such as phone number [3], International Mobile Subscriber Identity (IMSI) [4], International Mobile Equipment Identity (IMEI) [5], etc., to the mobile network service provider for authentication and access control. Therefore, a network service provider can obtain such information by parsing the authentication signaling traffic. However, smart devices that access the network through Wi-Fi are not authenticated to the mobile network service provider. Thus, it is not straightforward to obtain the mobile phone numbers with the traditional methods.

In this paper, we propose a human-centered smart framework for extraction of mobile phone number from home broadband network traffic. The proposed framework covers traffic collection, traffic data preprocessing, pattern matching and detail record output. Furthermore, we introduce a smart automated signature extraction method to extract mobile phone number from traffic to explore effective and low-cost way of data awareness for Human-Centered Smart Home Systems.

Traditional methods such as manual statistics, telephone inquiries, etc., are inefficient in practice [6]. Moreover, authenticity cannot be guaranteed with the traditional methods. With more and more Internet access through Wi-Fi, human-centered smart devices have become *underground traffic stations* for the spread and attacks of malicious information due to fluidity and concealment [7]. As a result, it is more difficult for the network security auditing and user traceability. In addition, operators cannot extract the phone number of a home subscriber to create a home profile. Such a profile can be further implemented into a precision convergence service for home users and marketing tools for inter-network users [8].

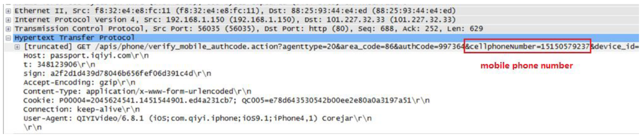
Traffic identification is important in traffic engineering, network security, network management and operation [9]. Port based, signature based, and statistical-feature based identification schemes are the mainstream approaches [10]–[12]. Among these methods, signature-based traffic identification is simple and relatively more efficient so that it is widely adopted by many network security systems [13]–[15]. A signature is a portion of payload data that is static and distinguishable for applications. It can be described as a sequence of strings or HEX values [16]. Recently, security of home network has attracted more attention due to the rapid development of smart home [17]. Some researchers have adopted the signature-based method to extract information from in-home Internet-of-Things (IoT) devices for security. Martin *et al.* [18] proposed a method called Pot2DPI which incorporates honey, Deep Packet Inspection (DPI) into home network security. However, when a protocol specification changes or a new protocol is adopted, network operators must start over to find valuable signatures. Therefore, automated signature extraction has received more attention from both research community and industry. Some researchers proposed automated signature extraction methods using string matching between normal and attack traffic for attacking. For example, Xu *et al.* [19] presented AutoSig which extracts multiple common substring sequences from sample flows as application signatures. However, there is limited research work related to the signature extraction of mobile phone number.

In this paper, a framework for extraction of mobile phone numbers from home broadband network traffic is proposed. In order to improve the efficiency and accuracy of detection, we further introduce an automated signature extraction scheme for mobile phone numbers in home Wi-Fi networks. The proposed method can achieve 86.2% accuracy in the real-life human-centered smart home network test. The rest of this paper is organized as follows. Section II presents the whole framework of extraction of mobile phone numbers from home network traffic. Section III describes the automated signature extraction method of mobile phone numbers from home network traffic. Section IV presents the experimental environment and the evaluation results. Section V concludes this work.

## II. FRAMEWORK OF THE MOBILE PHONE NUMBER EXTRACTION SCHEME

### *A. CHALLENGES OF THE INFORMATION EXTRACTION SCHEME*

A large number of mobile Internet applications (denoted as APPs hereafter) are registered with the mobile phone number for authentication. In addition, there are many mobile APPs that require the mobile phone numbers of users. For example, APPs for online shopping delivery, take-out delivery, etc., need the mobile phone numbers to provide further services. These APPs report information such as mobile phone numbers, IMSI, IMEI, etc., to cloud servers. It is feasible for a network service provider and cyber security agency to obtain such user information by parsing fixed broadband network traffic. For example, testing tools of Android, such as Monkey [20], can be used to trigger mobile APP events. A mobile phone number is usually exchanged in two ways: one is registered with the APP; the other one is obtained by manual confirmation, e.g., shipping confirmation during online shopping checkout. Fig. 1 shows an example of mobile phone number extraction by parsing Internet Protocol (IP) packets from network traffic.



```
⊞ Ethernet II, Src: f8:32:e4:e8:fc:11 (f8:32:e4:e8:fc:11), Dst: 88:25:93:44:e4:ed (88:25:93:44:e4:ed)
⊞ Internet Protocol Version 4, Src: 192.168.1.150 (192.168.1.150), Dst: 101.227.32.33 (101.227.32.33)
⊞ Transmission Control Protocol, Src Port: 56035 (56035), Dst Port: http (80), Seq: 688, Ack: 252, Len: 629
⊟ Hypertext Transfer Protocol
  ⊞ [truncated] GET /apis/phone/verify_mobile_authcode.action?agenttype=20&area_code=86&authCode=997364&cellphoneNumber=15150579237&device_id=
    Host: passport.iqiyi.com\r\n
    t: 348123906\r\n
    sign: a2f2d1d439d78046b656fef06d391c4d\r\n
    Accept-Encoding: gzip\r\n
    Content-Type: application/x-www-form-urlencoded\r\n
    Cookie: P00004=2045624541.1451544901.ed4a231cb7; QC005=e78d643530542b00ee2e80a0a3197a51\r\n
    Connection: keep-alive\r\n
    User-Agent: QIYIvideo/6.8.1 (iOS;com.qiyi.iphone;iOS9.1;iPhone4,1) Corejar\r\n
    \r\n
```

**FIGURE 1.** Example of parsing a mobile APP IP packet.

However, it is not always straightforward to profile a user, e.g., his/her mobile phone number, from data packets. There are a few challenges to be tackled, stated as follows.

- **Traffic data cleaning and filtering**: The number of users and collected user data can be categorized as big data, since they are massive in volume, frequent and huge in growth. Therefore, it is challenging to cleanse and filter the traffic data efficiently and accurately.
- **Encrypted traffic**: Some APPs use cryptographic mechanisms to encrypt data traffic. It is challenging to extract the mobile phone number in such APPs since parsing packets would be impossible [21].
- **Automated signature extraction**: Getting accurate keywords of packet signatures is the key to analyze and extract mobile phone number from network traffic. Traditionally, researchers or signature development engineers find signature keywords manually. It is challenging to find and update signature keywords efficiently and accurately. In the meantime, some numbers similar to a mobile phone number are also carried in IP packets. For example, a number sequence represents the identification of the user, or a service phone number of the APP, etc. Therefore, it is also challenging to find the signature keywords to judge the authenticity of the mobile phone number and family attribution.

### *B. AN OVERVIEW OF THE PROPOSED FRAMEWORK*

In this work, we propose a framework of mobile phone number extraction scheme to parse useful information from home network data traffic. An overview of the proposed framework is shown in Fig. 2. The framework comprises of *traffic collection*, *data pre-processing*, *signature matching* and *Smart Home Detail Record (SHDR)*.

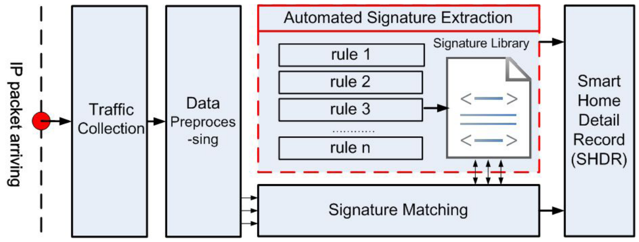


**FIGURE 2.** Framework of the extraction scheme.

*Traffic collection* is to collect the raw traffic from the core network link. There are two factors taken into consideration: one is traffic collection points, the other one is collection methods. Traffic collection points are usually located in different points of the communication network, such as the core network, aggregation layer, the broadband remote access server based access devices, etc. It can also be deployed inside the home gateway [22]. Collection methods often use the fiber splitter such as Test Access Point (TAP) [23], Traffic Measurement Agent (TMA) [24], Probe [25], etc., to copy the traffic in the network link, and then send to the server's network card. The server will complete the subsequent processing of the traffic.

*Data pre-processing* is to decode, cleanse and filter the collected traffic data since the collected traffic data includes a large amount of noisy data. After data pre-processing, the core function only processes a small portion of the collected data, which greatly reduce the computational overhead of the scheme. It is found that most mobile APPs use HTTP to communicate with the servers [19]. In the paper, we will use HTTP to demonstrate the proposed information extraction scheme, and other protocols will be our future research task. The collected HTTP traffic information includes a variety of encoded data, such as URL encoding [26], UTF-8 encoding [27], etc. Such data, especially URL information, contains a large amount of invalid or missing data, which requires different degrees of cleansing. In addition, some mobile APPs encrypt the data traffic. Therefore, proper decryption is required to perform signature matching. Fortunately, such encrypted data is not significant among the overall data traffic. For simplicity, analysis and identification of encrypted traffic will not be discussed in this paper and it will be explored in our future work. The granularity of data filtering will directly affect the efficiency of device information extraction, thereby affecting the real-time and accuracy of the entire system. Besides, filtering policies play an important role. For example, the policy can be based on user type or location, communication protocol, HTTP suffix, URL suffix, etc. In the proposed scheme, deep packet inspection (DPI) is used for filtering [15].

*Signature matching* is to match the traffic pattern from a signature library built by automated signature extraction to figure out the mobile phone number. The key is to constantly judge various signature keywords from the cleansed and filtered HTTP traffic data. Pattern-matching is the basic technique to realize this process, namely, signature string matching. Typically, signature strings are described using the Regular Expression standard syntax. A regular expression (RE) is a formal mathematical expression using a limited set of operators [28]. The purpose is to serve as concise specifications of sets of strings with certain desirable properties. For example, we assume $X$ is the set of letters $\{a, b, \ldots, z\}$, and the infinite set of all string over $X$ is written $Y$, that is:

$$Y = \{\varepsilon, a, b, \ldots, z, aa, ab, \ldots az, ba, bb, \ldots, bz, \ldots\}, \quad (1)$$

when $\varepsilon$ stands for the empty string. The syntax of REs can be compactly described by a generative grammar, such as:

$$E ::= a|\epsilon|E_1 + E_2|E_1E_2. \quad (2)$$

That is, the simplest REs consist of a single symbol $a$ from $X$ or the "unit expression" $\epsilon$. Smaller REs $E_1$, $E_2$ can be combined to form larger ones by the "sum operator" $E_1 + E_2$ or the "sequence opterator" $E_1E_2$. In this work, Hyperscan string matching algorithm is used to achieve signature matching based on automata rather than back-tracking. The finite automata is also called non-deterministic finite automata (NFA), due to the fact that there can be more than one state with pebbles on it when running it. This is done to distinguish them from the special case where exactly on state can ever have pebbles on it, in which case the automaton is called deterministic finite automata (DFA). Any NFA can be converted to a DFA. Hyperscan has the advantages of higher performance, smaller database, smaller stream state, etc. Traditional signature matching schemes require manual operation to form the signature library. However, they are time-consuming while inaccurate in some cases. Our proposed automated signature extraction scheme will be discussed in the next section.

*SHDR* is mainly used to describe smart devices in smart homes. For example, a smart device can be a mobile phone, a tablet, a PC, a TV Box, etc. The main information contained in SHDR includes timestamp, user IP, user broadband account, mobile phone number, IMSI, IMEI, source APP, host name, URL, Refer, UserAgent, smart device type, etc. SHDRs will be sent to the cloud platform for further data analytics.

## III. THE SMART AUTOMATED SIGNATURE EXTRACTION SCHEME

In this section, we propose a smart automated signature extraction scheme of mobile phone numbers from home network traffic. As shown in Fig. 3, once the incoming IP packets have been cleansed and filtered, they will pass through the extraction scheme by a set of rules. The matched signature will be the output of the scheme. The rules of automated signature extraction includes:

*Rule 1 (Classification of Smart Devices):* In most cases, a mobile phone number is only derived from one smart device. The type of the smart devices can be determined by using the UserAgent field in the HTTP GET packet firstly. The server uses UserAgent to identify the type of terminal,

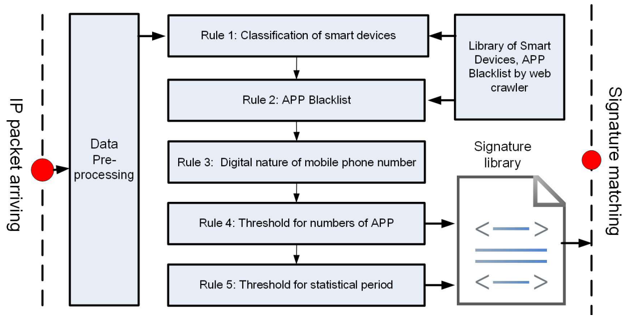

IP packet arriving
Data Pre-processing
Rule 1: Classification of smart devices
Rule 2: APP Blacklist
Rule 3: Digital nature of mobile phone number
Rule 4: Threshold for numbers of APP
Rule 5: Threshold for statistical period
Library of Smart Devices, APP Blacklist by web crawler
Signature library
Signature matching

**FIGURE 3.** The automated signature extraction scheme.

operating system and version, CPU type, browser and version, browser rendering engine, browser language, browser plug-in [29], [30], etc. By terminal matching, the terminal type can be accurately determined. When a terminal other than mobile phone is matched, it can be determined that the captured sequence of number is not a real mobile phone number.

*Rule 2 (APP Blacklist):* This rule is to blacklist some APPs from the analysis. With a long-term test, we found out that there are many APPs use mobile phone numbers. Among them, two types of information are similar to the mobile phone numbers of home users. One type includes some digital strings that may represent an identification, a serial number, etc. The other type includes some digital strings that are indeed mobile phone numbers but not of the home users. This type of phone numbers could be embedded in advertisements for business purposes, such as Yelp, or service numbers such as the AT&T customer service support. A great deal of noise data interference can be reduced by filtering out the mobile phone number information generated by such APPs. The host field information obtained after cleansing and filtering represents the host name of the packet where the mobile phone number keyword is located. It is also related to the name of the APP. Therefore, it is important to resolve the host name. In the proposed scheme, we use the webmagic [31] to crawl the relationship between a host name and an APP name in advance and encapsulate them as a matching function. The matching function can be used to determine which APP the phone number keyword is from accurately. Many different host names generated by the same APP can be merged into the same APP. The noise data interference from the APPs on the blacklist would be removed.

*Rule 3 (Digital Pattern of Mobile Phone Number):* As the mobile phone number often complies with a certain number of norms, which contains some digital patterns. For example, a mobile phone number of the China Mobile Communication has 11 digits, e.g., 139XXXXYYYY. The first three digits (i.e., 139) indicate the operator, the intermediate four digits (i.e., XXXX) represent the location of home register. As a result, these rules-based mobile phone numbers can be used to heuristically retrieve the keywords that are carried before the phone number. However, the signature of the mobile phone number extracted by this method is still not accurate because of two reasons. Firstly, the character string contained in the data packet that has the signature of the mobile phone number cannot be identified as the mobile phone number. It can be a series of numbers, such as an APP identification number which happens to meet the phone number specifications. Secondly, the phone number may not come from home users.

*Rule 4 (Threshold for Numbers of APPs):* The rules 1, 2 and 3 described above cannot guarantee the accuracy of the obtained mobile phone number information. This rule is a further verification based on the assumptions as follows: *the real phone number from home network traffic tends to be generated by more than one APP, i.e., COUNTapp* >1. Many APPs use the mobile phone numbers for authentications. When mobile phone users use this type of APPs at home, it is highly possible that data traffic carries the mobile phone number. In order to demonstrate the principle of the Rule 4, let us define some notations and terms related to the rule.

1) S: the Smart Home set, which is the set of all the smart homes under test, that is

$$S = \{s_1, s_2, s_3 \ldots s_n\}, \tag{3}$$

when $s$ refers to the smart home, $s \epsilon S$, $1 \leq n \leq N$, $N$ refers to the total numbers of smart homes under test. For example, $s_1$ refers to smart home1, $s_n$ refers to the $n$th smart home.

2) D: the set of smart mobile phone devices corresponding to some specified smart home, that is

$$D = \{\theta_1, \theta_2, \theta_3 \ldots \theta_k\}, \tag{4}$$

$\theta$ refers to the smart mobile phone device, such as iPhoneX. $1 \leq k \leq K$, $K$ is the maximum of smart mobile phone devices numbers of a smart home, which can be defined according to experience like 10 or more.

3) P: the phone numbers set of the smart mobile phone devices corresponding to some specified smart home, that is

$$P = \{\gamma_1, \gamma_2, \gamma_3 \ldots \gamma_j\}, \tag{5}$$

$1 \leq j \leq J$, $J$ is the total numbers of smart mobile phone numbers in a smart home. Apparently, $\gamma$ is corresponding to $\theta$ in the set of $D$.

4) A: the set of mobile application like facebook or twitter, that is

$$A = \{\alpha_1, \alpha_2, \alpha_3 \ldots \alpha_i\}, \tag{6}$$

$1 \leq i \leq I$, $\alpha$ refers to the mobile application of smart mobile phone. $I$ is the number of the mobile applications.

According to above mentioned, we discuss the rule 4, especially the threshold of the numbers of Apps. For a specified smart home $s_n$, $s_n \epsilon S$, each $\theta_k$, $\theta_k \epsilon D$, has 1 or more mobile phone numbers, which is represented by $\gamma_j$, $\gamma_j \epsilon P$. According to above mentioned about the scheme of extraction, each phone number, that is $\gamma_j$ is extracted from the traffic generated by a mobile App which is represented as $\alpha_i$. Note that a mobile phone number can be extracted from multiple mobile Apps, that is to say, a $\gamma_j$ can be corresponding to multiple $\alpha_i$.

In order to improve the accuracy of the extraction of mobile phone numbers, it is beneficial to optimizing the algorithm that defining the threshold of numbers of Mobile Apps where the mobile phone numbers can be extracted. As we all know, it is not enough to decide that the mobile phone number extracted from some Mobile Apps is the right one if this phone number can be found in only a few Mobile Apps. There are some coincidence, for example, your neighbors or friends visit your home and use their mobile phone accessed by your home network. Therefore, it is very important to define the threshold of numbers of Mobile Apps where the mobile phone numbers can be extracted, here we define *Threshold_A* is the threshold. Usually, set the *Threshold_A* 10 or more would be better for the algorithm.

*Rule 5 (Threshold for Statistical Period):* This rule is similar to rule 4. It is based on the assumption as follows: *noise data that comply with mobile phone number specifications are often temporary, while traffic data carrying real cell phone numbers is stable and normal*. A packet that carries a real phone number should appear multiple times. Therefore, the number of days that traffic data carrying a mobile phone number appears may be used as a metric for the accuracy of the mobile phone number extraction.

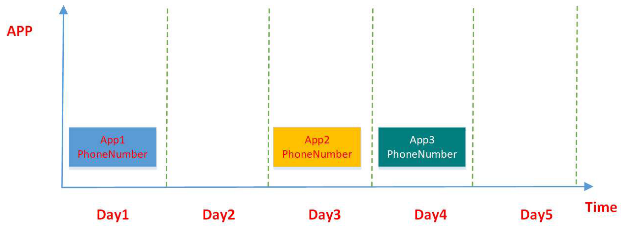


**FIGURE 4.** The occurrence of the same mobile phone number in different day from different mobile Apps.

There will be two cases in real-life applications. As shown in Fig. 4, one case is that the flow of the same mobile phone number data can be generated by different APPs on different days. As shown in Fig. 5, the other case is that the traffic data of the same mobile phone number can be generated by different APPs on the same day. The time distribution in the two figures are different. Therefore, the traffic data of the mobile phone number generated by the APPs on different days cannot be used as a metric to determine the accuracy of the mobile phone number extraction. In addition, the impact of Cookies also need to be considered. Because many applications cache to local data files or cookies the user information, including phone number information, when they first use the APP. The data traffic will not carry the user information in the later transmissions unless there is a data file or cookie removal operation message from the user. This results in a smaller number of days when the user's mobile number information appears.

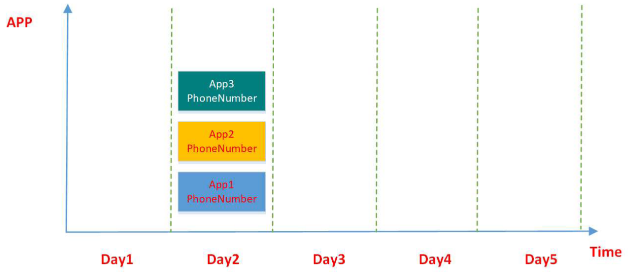


**FIGURE 5.** The occurrence of the same mobile phone number in same day from different mobile Apps.

Above all, the threshold for numbers of APP (i.e., rule 4) is taken as the main basis for determining the accuracy of the phone number. The threshold for statistical period (i.e., rule 5) is taken as a supplementary basis.

In summary, the method for judging the accuracy of mobile phone numbers is described as follows:

1) Through the smart devices matching rule, non smart phone traffic data is filtered out.
2) Through the APP blacklist rule, non home users' mobile number traffic data is filtered out.
3) Through the signature rule of the mobile phone number, the traffic data containing the digital characteristics of the mobile phone number is extracted and a quaternion (i.e., {Home Broadband Account, phone number, APP source, timestamp}) is formed.
4) According to the rule of threshold for numbers of APP, the pair {Home Broadband Account, phone number} is taken as the key to calculate the number of APPs appearing in the traffic. If the number of APPs exceeds the threshold set by the system, the phone number is considered as the real home phone number, otherwise it will be further checked with the next step.
5) According to the rule of threshold for statistical period, the pair {Home Broadband Account, phone number} is taken as the key to calculate the number of days that it appears. If the number of days exceeds the threshold set by the system, it is considered as the real home phone number.
6) Finally, the key string in front of the mobile phone number and the APP of the mobile phone are written as signatures into the signature library. At this time, the construction of the signature library is completed.

An example of the signature library is shown in Table 1. *Key string* is the corresponding signature string of mobile phone number in a HTTP message packet. *APP identity* represents the traffic originator. For example, "10086" is the APP identity of China Mobile Communication. Consequently, we can extract mobile phone number of home users from traffic by using DPI methods such as string matching algorithm with the signature library.

**TABLE 1.** An example of the signature library.

| Key string | APP identity |
|---|---|
| ?tel= | 10086 |
| &mobile= | 10086 |
| &phoneno= | Android Market |
| ?mobile= | Kuwo music box |
| ?passport= | Lianzhong hall |
| &phone= | Sina Weibo |

## IV. EXPERIMENT AND EVALUATION RESULTS

In this section, we present the experiment and evaluation results of the proposed scheme based on real-life data.

### A. NETWORK SETTINGS OF THE EXPERIMENTS

The statistical information of the experiments are shown in Table 2. The experimental environment is set as a human-centered smart home network, which covers 1.2 million broadband home users.

**TABLE 2. Statistical information of the experiments.**

| Items | Statistics |
|---|---|
| number of home broadband users | 1.2 million |
| number of smart devices | 1.5 million |
| number of smart devices brands | 294 |
| number of smart devices models | 6500 |
| number of original data records | 90.72 million |
| number of filtered data records | 18.14 million |

The traffic collection point is located beside the BRAS equipment. By mirroring the upstream traffic of the BRAS, the traffic collection server receives the traffic data, performs data cleaning, filtering, and then sends the traffic to the Hadoop platform. The Hive performs the automated signature extraction method of mobile phone numbers to form a signature library. At the same time, the Hadoop platform performs high-speed signature matching and outputs the mobile phone number list in SHDR format. An overview of the network setting is shown in Fig. 6.

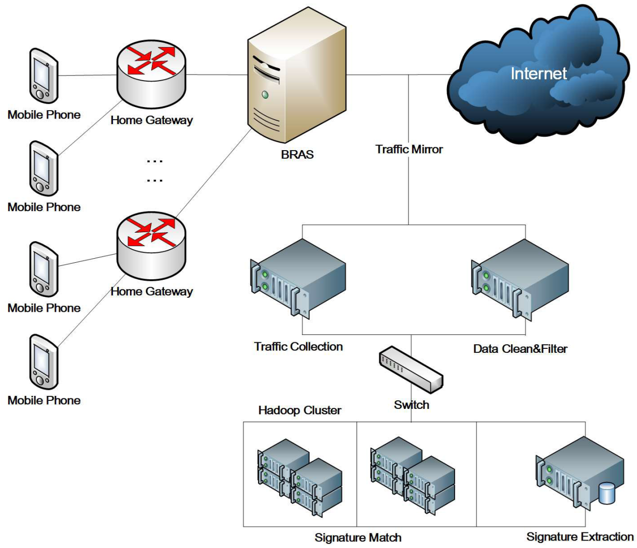


**FIGURE 6. Network setup of the experiments.**

### B. DATA ANALYSIS

We first show the results regarding the amount and accuracy of mobile phone numbers extracted from home network traffic. Assume that each mobile phone number only belongs to one individual home broadband account assigned by the network service provider.

As shown in Fig. 7, about 1 million home broadband accounts have been successfully decoded from the data traffic. The rest 0.2 million users were missing due to a few reasons. For example, no authorization packet was sent to the accounting server because of the missing authentication process from a home gateway. It may also because that the home broadband is off-line. Among the 1 million accounts decoded, roughly 200,000 mobile phone numbers can be extracted precisely, with the successful ratio of matched mobile phone numbers with home broadband accounts above 20%. Moreover, number of decoded home broadband accounts and the number of extracted phone numbers increased to 1.1 million and 90% with an observation period of 7 days. Therefore, we can see that the amount of mobile phone numbers with matching the home broadband account are steadily and gradually increasing over time.

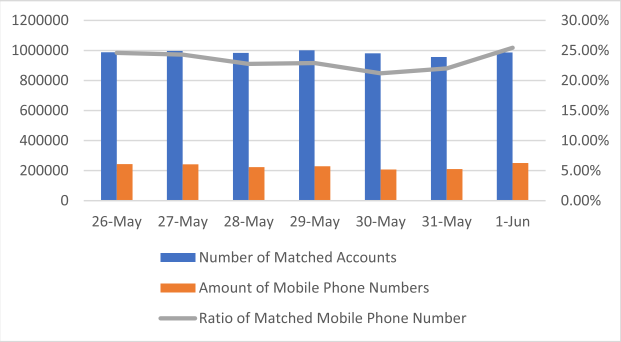


**FIGURE 7. Matching analysis between home broadband account and mobile phone number.**

We then show the comparison between local mobile phone numbers and the roaming ones. As shown in Fig. 8, the ratio is about 8:2. Such information can be further applied to many applications. For example, it may represent the ratio of non-residents in the current area, etc.

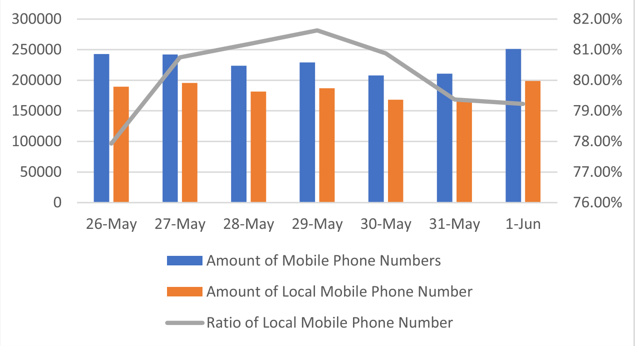


**FIGURE 8. Comparative analysis of local and other cities phone numbers.**

### C. PERFORMANCE EVALUATION OF THE AUTOMATED SIGNATURE EXTRACTION SCHEME

#### 1) PERFORMANCE METRICS

The performance of the automated signature extraction scheme is evaluated in terms of *accuracy and identification error rate (IER)*. $TruePositives(TP)$ is the number of correctly extracted mobile phone numbers, $FalsePositives(FP)$ is the number of mobile phone numbers falsely extracted.

- ***accuracy***: the *rate of accuracy* is the ratio of the number of correct phone number to the total number of all access phone number in the test environment. It can be calculated as the ratio of the sum of all *TP* to the sum of all the *TP* and *FP*, that is

$$accuracy = \frac{TP}{TP + FP}. \tag{7}$$

- ***identification error rate (IER)***: The *identification error rate* is the ratio of the number of phone number falsely extracted to the total number of all access phone number in the test environment. We can calculate it as the ratio of *FP* over the sum of *TP* and *FP*, that is

$$IER = \frac{FP}{TP + FP}. \tag{8}$$

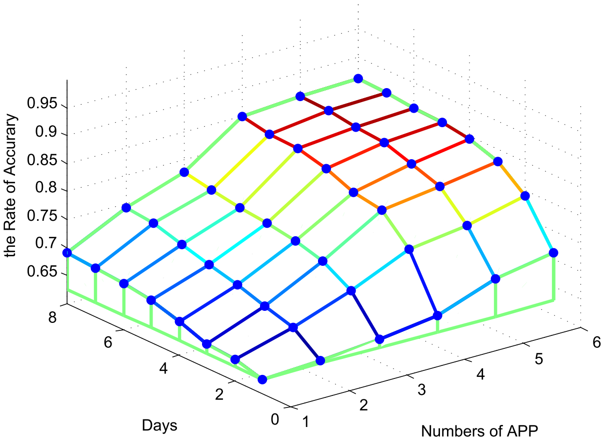


**FIGURE 9. Accuracy of the proposed scheme.**

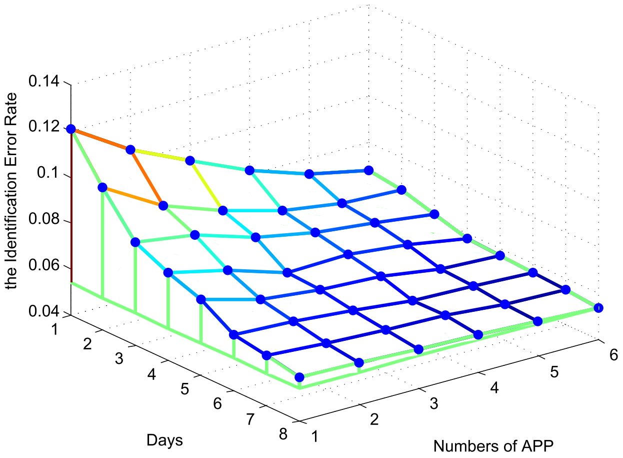


**FIGURE 10. Identification error rate.**

### 2) PERFORMANCE ANALYSIS

As shown in Fig. 9, with the increase of threshold for numbers of APP and threshold for statistic period, the accuracy is obviously improved. In this experimental environment, the accuracy is up to 86.2% when the threshold for numbers of APP is 6 and the threshold for statistical period is 8 days.

As shown in Fig. 10, with the increase of threshold for numbers of APP and threshold for statistic period, identification error rate obviously decreases. In the experimental environment, when the threshold for numbers of APP is 6 and the threshold for statistical period is 8 days, the identification error rate can be controlled as low as 5.4%.

## V. CONCLUSION

In the paper, we proposed a smart framework for profiling users by extracting the mobile phone numbers. The proposed framework is based on home broadband data traffic analysis and DPI. In order to counter the inefficiency and inaccuracy of traditional DPI methods, we further proposed an automated signature extraction scheme. The real-life broadband network test demonstrated that our proposed scheme can achieve an accuracy up to 86.2%. In the future work, we will continue to improve the accuracy of the scheme and expand the study of human-centered smart home profiling with other user signatures.

## REFERENCES

[1] J. Ji *et al.*, "A human-centered smart home system with wearable-sensor behavior analysis," in *Proc. IEEE Int. Conf. Automat. Sci. Eng. (CASE)*, Aug. 2016, pp. 1112–1117.

[2] F. K. Santoso and N. C. H. Vun, "Securing IoT for smart home system," in *Proc. Int. Symp. Consum. Electron. (ISCE)*, Jun. 2015, pp. 1–2.

[3] S. Jiang, B. Wei, T. Wang, Z. Zhao, and X. Zhang, "Big data enabled user behavior characteristics in mobile Internet," in *Proc. 9th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2017, pp. 1–5.

[4] R. V. Bulychev, D. E. Goncharov, and I. F. Babalova, "Obtaining IMSI by software-defined radio (RTL-SDR)," in *Proc. IEEE Conf. Russian Young Res. Electr. Electron. Eng. (EIConRus)*, Jan./Feb. 2018, pp. 21–23.

[5] A. J. F. Loureiro, D. Gallegos, and G. Caldwell, "Substandard cell phones: Impact on network quality and a new method to identify an unlicensed IMEI in the network," *IEEE Commun. Mag.*, vol. 52, no. 3, pp. 90–96, Mar. 2014.

[6] *How to Monitor, Measure, and Manager Your Broadband Consumption*. Accessed: Apr. 6, 2017. [Online]. Available: https://www.pcworld.com/article/3072638/home-networking/how-to-measure-monitor-and-manage-your-broadband-consumption.html/

[7] D. Vavilov, A. Melezhik, and I. Platonov, "Reference model for Smart Home user behavior analysis software module," in *Proc. IEEE 4th Int. Conf. Consum. Electron.–Berlin (ICCE-Berlin)*, Sep. 2014, pp. 3–6.

[8] J. L. Garcia-Dorado, A. Finamore, M. Mellia, M. Meo, and M. Munafo, "Characterization of ISP traffic: Trends, user habits, and access technology impact," *IEEE Trans. Netw. Service Manag.*, vol. 9, no. 2, pp. 142–155, Jun. 2012.

[9] W. de Donato, A. Pescapé, and A. Dainotti, "Traffic identification engine: An open platform for traffic classification," *IEEE Netw.*, vol. 28, no. 2, pp. 56–64, Mar. 2014.

[10] F. Hajikarami, M. Berenjkoub, and M. H. Manshaei, "A modular two-layer system for accurate and fast traffic classification," in *Proc. 11th Int. ISC Conf. Inf. Secur. Cryptol.*, Sep. 2014, pp. 149–154.

[11] P. Lizhi, Y. Bo, C. Yuehui, and W. Tong, "How many packets are most effective for early stage traffic identification: An experimental study," *China Commun.*, vol. 11, no. 9, pp. 183–193, Sep. 2014.

[12] T. Iwai and A. Nakao, "Adaptive mobile application identification through in-network machine learning," in *Proc. 18th Asia–Pacific Netw. Oper. Manage. Symp. (APNOMS)*, Oct. 2016, pp. 1–6.

[13] H.-J. Kang, M.-S. Kim, and J. W.-K. Hong, "A Method on Multimedia Service Traffic Monitoring and Analysis," in *Proc. Int. Workshop Distrib. Syst., Oper. Manage.*. Berlin, Germany: Springer, 2003, pp. 93–105, doi: 10.1007/978-3-540-39671-0_9.

[14] J. van der Merwe, R. Cáceres, Y.-H. Chu, and C. Sreenan, "mmdump: A tool for monitoring Internet multimedia traffic," *SIGCOMM Comput. Commun. Rev.*, vol. 30, no. 5, pp. 48–59, 2000, doi: 10.1145/505672.505678.

[15] S. Sen, O. Spatscheck, and D. Wang, "Accurate, scalable in-network identification of p2p traffic using application signatures," in *Proc. 13th Int. Conf. World Wide Web*. New York, NY, USA: ACM, 2004, pp. 512–521, doi: 10.1145/988672.988742.

[16] B.-C. Park, Y.-J. Won, M.-S. Kim, and J. W. Hong, "Towards automated application signature generation for traffic identification," in *Proc. IEEE Netw. Oper. Manage. Symp. (NOMS)*, Apr. 2008, pp. 160–167.

[17] E. Fernandes, J. Jung, and A. Prakash, "Security analysis of emerging smart home applications," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2016, pp. 636–654.

[18] V. Martin, Q. Cao, and T. Benson, "Fending off IoT-hunting attacks at home networks," in *Proc. 2nd Workshop Cloud-Assist. Netw.* New York, NY, USA: ACM, 2017, pp. 67–72. [Online]. Available: http://doi.acm.org.libproxy.udayton.edu/10.1145/3155921.3160640

[19] Q. Xu *et al.*, "Automatic generation of mobile app signatures from traffic observations," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr./May 2015, pp. 1481–1489.

[20] M. K. Alzaylaee, S. Y. Yerima, and S. Sezer, "Improving dynamic analysis of Android apps using hybrid test input generation," in *Proc. Int. Conf. Cyber Secur. Protection Digit. Services (Cyber Secur.)*, Jun. 2017, pp. 1–8.

[21] X. Chen, P. Wang, and S. Liu, "Key technology research of SSL encrypted application identification under imbalance of application class," *Telecommun. Sci.*, vol. 31, no. 12, p. 355, 2015.

[22] P. Wang and X. Chen, "Co_Hijacking monitor: Collaborative detecting and locating mechanism for HTTP spectral hijacking," in *Proc. IEEE 15th Int. Conf. Dependable, Autonomic Secure Comput., 15th Int. Conf. Pervasive Intell. Comput., 3rd Int. Conf. Big Data Intell. Comput. Cyber Sci. Technol. Congr. (DASC/PiCom/DataCom/CyberSciTech)*, Nov. 2017, pp. 61–67

[23] *Network Tap*. Accessed: Mar. 13, 2017. [Online]. Available: https://en.wikipedia.org/wiki/Network_tap

[24] A. Kozemcak and T. Kovacik, "Different network traffic measurement techniques—Possibilities and results," in *Proc. ELMAR*, Sep. 2012, pp. 93–96.

[25] R. Hofstede *et al.*, "Flow monitoring explained: From packet capture to data analysis with netflow and IPFIX," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 2037–2064, 4th Quart., 2014.

[26] K. Cheng, R. Guo, and M. Gao, "An optimizing chinese string matching algorithm based on the URL encoding," in *Proc. WASE Int. Conf. Inf. Eng.*, Aug. 2010, pp. 23–25.

[27] Z. Qiang and W. Qiang, "The processing method of Chinese characters in TCL scripting language," in *Proc. 1st ACIS Int. Symp. Cryptogr., Netw. Secur., Data Mining Knowl. Discovery, E-Commerce Appl., Embedded Syst.*, Oct. 2010, pp. 398–400.

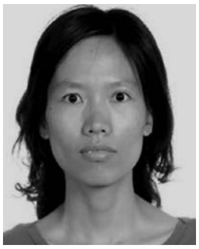
[28] C. Xu, J. Su, S. Chen, and B. Han, "Offset-FA: Detach the closures and countings for efficient regular expression matching," in *Proc. IEEE 7th Int. Symp. Cloud Service Comput. (SC2)*, Nov. 2017, pp. 263–266.

[29] P. Wang, "Big data plug-in technology for smart router based on multidimensional awareness," *J. Nanjing Univ. Posts Telecommun. (Natural Sci. Ed.)*, vol. 36, no. 4, pp. 18–21, 2016.

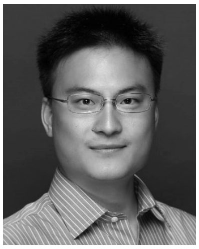
[30] P. Wang, F. Ye, and X. Chen, "Smart devices information extraction in home Wi-Fi networks," *Internet Technol. Lett.*, vol. 1, no. 3, p. e42, 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/itl2.42

[31] *WebMagic in Action*. Accessed: Aug. 21, 2017. [Online]. Available: http://webmagic.io/docs/zh/

**PAN WANG** (M'18) received the B.S. degree from the Department of Communication Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China, in 2001, and the Ph.D. degree in electrical and computer engineering from the Nanjing University of Posts and Telecommunications, in 2013. From 2017 to 2018, he was a Visiting Scholar with the Department of Electrical and Computer Engineering, University of Dayton. He is currently an Associate Professor with the School of Modern Posts, Nanjing University of Posts and Telecommunications. His research interests include cyber security and communication network security, network measurements, quality of service, deep packet inspection, SDN, and big data analytics and applications.



**XUEJIAO CHEN** (M'18) received the B.S. degree from the Department of Communication Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China, in 2001, and the master's degree in electrical and computer engineering from the Nanjing University of Posts and Telecommunications, in 2006. In 2017, she was a Visiting Scholar with the Department of Electrical and Computer Engineering, University of Dayton. She is currently an Assistant Professor with the Department of Communication Engineering, Nanjing College of Information Technology, Nanjing. Her research interests include wireless communications and networks, cyber security and communication network security, network measurements, quality of service, and deep packet inspection.



**FENG YE** (S'12–M'15) received the B.S. degree from the Department of Electronics Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2011, and the Ph.D. degree in electrical and computer engineering from the University of Nebraska–Lincoln (UNL), Lincoln, NE, USA, in 2015. He was with the Department of ECE, UNL, as an Instructor and a Researcher, from 2015 to 2016. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, University of Dayton, Dayton, OH, USA. His research interests include cyber security and communication network security, wireless communications and networks, green ICT, smart grid communications and energy optimization, and big data analytics and applications. He serves as a TPC member for numerous international conferences, including INFOCOM, GLOBECOM, VTC, and ICC. He was a recipient of the 2015 Top Reviewer of the IEEE Vehicular Technology Society. He serves as the Co-Chair of ICNC'19 Signal Processing for Communications Symposium, the Publicity Co-Chair of the IEEE CBDCom 2018, the Co-Chair of Cognitive Radio and Networking Symposium and the IEEE ICC 2018. He is currently an Associate Editor of *Security and Privacy* (Wiley), and *China Communications*. He is also a reviewer for several IEEE journals, including the IEEE TRANSACTIONS ON BIG DATA, the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING, the IEEE TRANSACTIONS ON SMART GRID, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He serves as the Secretary of the IEEE Technical Committee on Green Communications and Computing.

**ZHIXIN SUN** was born in Xuancheng, China, in 1964. He received the Ph.D. degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 1998. From 2001 to 2002, he held a post-doctoral position with the School of Engineering, Seoul National University, South Korea. He is currently a Professor and the Dean of the School of Modern Posts, Nanjing University of Posts and Telecommunications. His research interests are in cloud computing, cryptography, and traffic identification.