Hindawi
Security and Communication Networks
Volume 2018, Article ID 2506319, 12 pages
https://doi.org/10.1155/2018/2506319

WILEY | Hindawi

Research Article

Behaviors of High-Frequency Subscribers in Cellular Data Networks

Jingtao Li,[1,2] **Yang Liu**,[1] **Wengang Pei**,[1] **and Zhen Cao**[1]

[1]*Software School, Fudan University, Shanghai, China*
[2]*Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China*

Correspondence should be addressed to Jingtao Li; lijt@fudan.edu.cn

Received 24 April 2018; Revised 10 September 2018; Accepted 20 September 2018; Published 6 November 2018

Academic Editor: Dimitrios Geneiatakis

Cellular networks operate under restrictive constraints of resources including radio channel capacity and network processing capability. The tremendous growth in the cellular data network usage brings operators with unprecedented signaling overloads and threatens the stability of the network. High-frequency subscribers, who produce low data volume but cause high signaling overhead, are extremely resource-inefficient. For subscribers who activate more than 5 sessions per hour, they only account for 1.19% of the total subscribers and generate about 3.81% data traffic but consume roughly 19.46% of the signaling resources, resulting in the inconsistent signaling-data bandwidth consumptions. Understanding the characteristics of those users has an important significance of capacity design and optimal allocation of resources. A lack of understanding of this active group potentially leads to low network performance and security threats. In this paper, we perform the first city-wide, large-scale investigation of high-frequency subscribers. By applying a set of novel approaches, such as pattern extraction and user behavior rebuilding, we observed that high-frequency subscribers correspond to a lower percentage of none-pattern traffic, showing positive correlation between access regularity and session activation frequency. Besides, we found that amount of high-frequency subscribers has abnormal behaviors, resulting in unwanted signaling loads. We demonstrate that our findings have significant implications on network optimization.

1. Introduction

The Internet is going wireless and mobile, and cellular network is going to be the favorite way to access network. However, in spite of a concerted effort to support packet-switched traffic, cellular data networks are still, at their essence, circuit-switched systems. Because of this inflexibility, the tremendous growth in the cellular data network usage brings operators with unprecedented signaling overloads and threatens the stability of the network.

High-frequency subscribers, one particular type of subscribers who access cellular network frequently and do be more active than others, are extremely resource-inefficient for high signaling overhead and low data transmission volume, resulting in the following potential threats. Firstly, high-frequency subscribers produce few data traffic but have disproportionately high signaling overhead. Secondly, high signaling resource consumption with few data transfers is unfair to other subscribers, which puts high signaling pressure on network operators but produces low fees.

The problem caused by those users is not a simple problem of user behavior but an all-round problem which affect network management, security, performance, and so on. Therefore, it is important to characterize their behaviors to balance resource usage and guarantee network performance. Besides, characterizing their behaviors can fill a vacancy in the analysis of such users on the one hand and deepen our understanding of real-world traffic in cellular networks on the other.

In this paper, based on the real-world traces collected from a commercial cellular network in China in 2010, we present the first in-depth city-wide measurement of high-frequency subscribers to quantitatively understand the following important characteristics from data session (data session: a period of continuous activity, lasting from the allocation to the release of network resource and

containing control-plane and data-plane messages) level and application-level semantics [1]:

(i) The impact on signaling resources of high-frequency subscriber sessions for commercial cellular networks.

(ii) The regularity of subscribers' high session activation behaviors.

(iii) The correlation between session activation frequency and periodicity.

(iv) The correlation between periodicity and abnormal behaviors

In summary, we detail our key phenomena as follows:

(i) **Inconsistent signaling-data bandwidth consumptions:** producing low data volume, but causing heavy signaling load, high-frequency subscribers result in the unfairness of resource allocation and loss of network operators. Furthermore, this may decrease the network performance.

(ii) **Positive correlation between regularity and activation frequency:** higher frequency corresponds to a lower percentage of none-pattern traffic. More than 40% of the subscribers ($N_{sa} \geq 10$) are in accordance with Pattern A (Pattern A: activate sessions per fixed seconds), and for those with N_{sa} larger than 20, almost 70 % of them are Pattern A subscribers.

(iii) **Amount of periodical subscribers actually does abnormal behaviors:** by correlating cross-layer info, we identify that nearly half of Pattern A users' behaviors are abnormal, the top four types of which are Periodical PDP Context Activation, Network-side Termination, Periodical UDP Packets, and Privacy Information Uploading. Producing low data volume, but doing some other bad things sneakily, such as heavy signaling load and privacy leak, those abnormal behaviors endanger the security and decrease network performance.

Paper organization. Section 2 provides the background and dataset. Section 3 points our research intention and methodology. Then we characterize high-frequency subscribers and three kinds of association in next three sections: quantitative relationship between data transfer and signaling consumption in Section 4, correlation between frequency and periodicity in Section 5, and relevance between periodicity and abnormal behaviors in Section 6. Next, we propose some real-life implications of our findings in Section 7. Section 8 summarizes related work and Section 9 concludes the paper.

2. UMTS and Dataset

Our traces are collected from Gn (Gn: the Data transmission link between SGSN and GGSN in cellular network) interface, in a cellular operator's core UMTS network which services a large metropolitan with a population of more than one millions in China. Figure 1 plots its architecture. We collected all two types of data traffic: one is **PDP control messages** between SGSN and GGSN for the initiation, termination and updating of data session, and another is **Tunneled IP packets** between mobile terminals and the GGSN. Our collection lasts for three days in January, 2010.

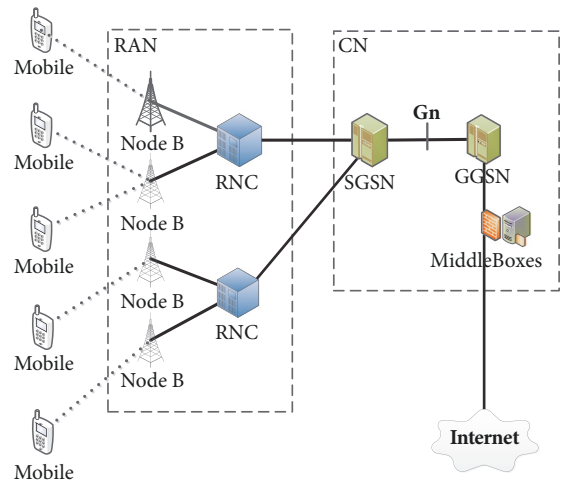

Figure 1: UMTS network architecture.

In our work, each data session is logged as a behavior record which is the reflection of a subscriber's continuous activities and is indexed by a session activation timestamp and an anonymized subscriber identifier (IMSI (IMSI: International Mobile Subscriber Identification Number)). Each record contains all kinds of key information: **control messages** which can reconstruct the signaling exchanges during the session, such as the messages for establishing and terminating the session, **IP packets** during the holding time of the session for key fields extraction, such as IP address, ports and compressed payload, and **application identification** such as port information, payload signatures, and other heuristics which are detailed in [2].

3. Methodology

3.1. Phenomenon Driven. High-frequency subscribers are extremely resource-inefficient for two phenomena: **low data volume** and **high signaling overhead**; i.e., subscribers ($N_{sa} \geq 5$) generate about 3.81% data traffic but roughly consume 19.46% of the signaling resources.

From network operator's perspective, they should apply much more policies to deal with this huge inconsistency, resulting in heavy management and poor network performance.

High-frequency network access is not a simple specific user behavior but an all-round complex problem which affect network management, security, performance, and so on. Therefore, we struggle to dissect those specific subscribers.

Unfortunately, by focusing on session-level features we observed that the inconsistency phenomenon between data volume and signaling consumption covers the entire spectrum of the frequency. To further dissect this inconsistency, we then correlate frequency with access regularity from user access level and consider the very/extremely high-frequency

subscribers to extract their behavior pieces from application-level semantics.

3.2. Why Session Centric? Different form prior work which analyzed based on packet or flow level or used an idle time (e.g., 5 minutes) to approximate the termination of a session, we studied from session perspective for the following considerations:

(i) Data session maintenance is the basic goal of network signaling while flow is just one piece of a session. A data session not only contains its flows' total information, but also includes some additional information such as session activation time.

(ii) To some extent, data session can be used to capture the resource usage behaviors of mobile subscribers by semantic analysis while flow cannot do.

(iii) One of our purposes is to extract high-frequency users' characteristics comprehensively and quantify the signaling consumption, and those information is shown in data session level brightly and perfectly.

Though it is much harder to rebuild sessions than flows, it is more useful to macroscopically analyze session signaling consumption than the analysis of flow pieces. In this paper, we reconstruct more than 910K complete data sessions and develop novel pattern-identify method to study their characteristics.

3.3. How to Extract Sessions. As mentioned above, different from prior work which terminated session based on an long idle time without data transfer, we accurately extract one data session from raw packet traces according to the following steps:

(i) Extract session control message including PDP Context Create, Update, and Delete messages, and rebuild session structure based on TEID_Control (Tunnel Endpoint Identifier for control-plane massages) option.

(ii) Extract session data message and join with session structure based on TEID_Data field, which identifies the data communication tunnel.

(iii) For one particular subscriber, correlate sessions based on IMSI option as IMSI is identifier that identify the unique data sessions activated by the same subscriber.

The methods proposed in [3, 4] extract sessions relied only on IP packets in data plane; that is, they assume the termination of a session if no packets from the private IP address of that session are observed for a threshold time. They mainly focus on how to estimate the signaling overheads brought by RRC state promotions and demotions in a single session, which are caused by the intervals between continuous IP packets. However, by correlating control plane and data plane, we can accurately identify the beginning and the end of a session. In addition, we can connect all the sessions belonging to one subscriber together, which will be used to analyze subscribers' session-level behaviors. Note that private IP addresses are dynamically allocated and normally one subscriber would get different IP addresses when connecting to the network twice, so methods based on IP addresses cannot be used to connect sessions.

3.4. Cross-Layer Analysis. To deeply understand high-frequency subscribers' behaviors, we should do know what they have done, how they have performed, and when and where they have communicated. However, those user information is hidden in various layers, such as access time info hidden in session layer, transport protocol info hidden in flow layer, and application preference info in application layer.

To restore a real comprehensive user from disorganized data, we should accurately extract fragments from various perspectives, correlate them together, and quantitatively analyze subscribers' behaviors.

In this paper, we firstly try to quantify the influence of this inconsistency phenomenon between data volume and signaling consumption which covers the entire spectrum of the frequency by focusing on session-level features. Then we step into user access level to correlate frequency with access regularity; finally we consider the very/extremely high-frequency subscribers to extract behavior pieces from application-level semantics.

4. Inconsistent Signaling-Data Bandwidth Consumptions

In this section, by quantifying the signaling resource impact of high-frequency subscribers measured, we intend to understand user behavior and quantify the correlation between traffic volume and signaling consumption.

4.1. The Phenomenon. We use two indicators for session clustering through most of our following analysis: (a) N_{sa}: defined as the average session-activate times per hour for every subscriber; (b) **SAI** (abbreviation for **S**ession **A**ctivation **I**nterval): defined as the interval between two adjacent session activation requests initiated by the same subscribers.

Figure 2 plots the CDFs of sessions and subscribers over the metrics. We observed that major subscribers (93%) have a N_{sa} less than 3, but just accounts for roughly 70% of the total sessions, suggesting that others (7%) have a high-frequency behavior to activate sessions, counting for approximately 30%. Furthermore, we observed that the top 1% of the subscribers create nearly 20% of the total data sessions, suggesting that a few subscribers activate far more sessions than others. This shows a significant imbalance of network usage among subscribers with a few subscribers hogging the much of the network resource, resulting in the unfairness of resources sharing.

4.2. Quantitative Analysis. Given the subscriber set "U", we use two metrics to profile its characteristics:(a) **traffic volume** (labeled as **V**), defined as the number of bytes above transport layer consumed by all subscribers; (b) **signaling overhead** (labeled as **S**), defined as the total signaling messages involved in creating and deleting sessions by all subscribers and containing two parts: one part is the signaling overheads for radio resource control, and the number of signaling messages

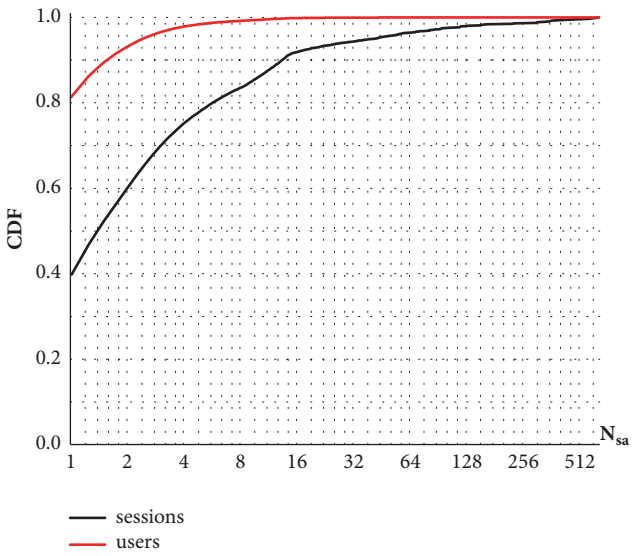

Figure 2: CDF of subscribers and sessions over N_{sa}.

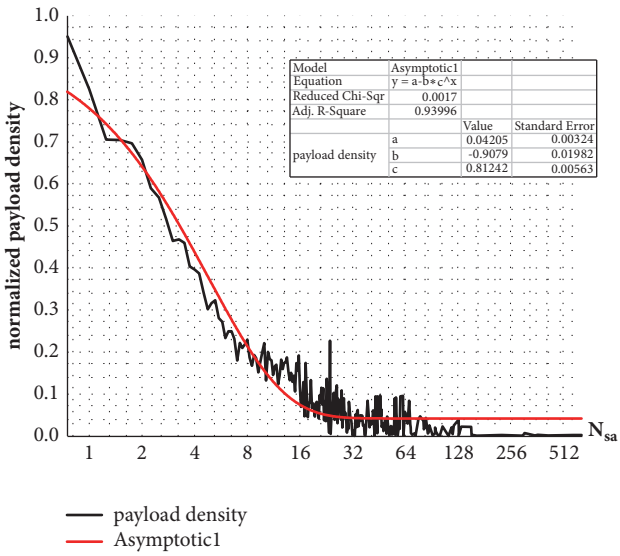

Figure 3: Payload density distribution.

is estimated based on the signaling exchanges by the RRC state transitions [3]; the other part is the signaling overheads of PDP context control, and the number of messages in this part are counted directly from our dataset (and this part of signaling overheads are not taken into consideration by the previous works [3, 4]).

As showed in Table 1, we totally focus on five sets: U_0 is the all subscribers in the entire dataset; and U_1 to U_4 correspond to the subscribers with N_{sa} larger than 5, 10, 15, and 20, respectively. And for subscriber set U_i, $V(U_i)$ (or $S(U_i)$) represents the total traffic volume (signaling overheads) consumed by these subscribers as defined above.

For each set U_i, we first calculated its subscriber subset UP_i detected in the real trace, which is the subscribers in U_i who have periodic session activation behaviors. Then we calculated the ratio of the traffic volume of U_i over that of U0 and the ratio of the traffic volume of UP_i over that of U_0 as

$$\Delta V\,(all) = \frac{V\,(U_i)}{V\,(U_0)}$$
$$\text{and } \Delta V\,(Periodicity) = \frac{V\,(UP_i)}{V\,(U_0)} \quad (1)$$

The ratio of the signaling overheads brought by sessions of U_i over those brought by sessions of U_0 and the ratio of the signaling overheads brought by sessions of UP_i over those brought by sessions of U_0 are also calculated as

$$\Delta S\,(all) = \frac{S\,(U_i)}{S\,(U_0)}$$
$$\text{and } \Delta S\,(Periodicity) = \frac{S\,(UP_i)}{S\,(U_0)}. \quad (2)$$

We observed that as N_{sa} grows larger, the traffic volume of high-frequency subscribers accounts for less proportion. But by removing these subscribers, the signaling overheads will reduce a lot. This trend is more obvious when applying it to periodic activated sessions. Clearly, there exists tremendous disparity between the traffic volume and the resource consumption of high frequent or periodical session activations, indicating that high frequent or periodical session activations are extremely resource inefficient.

As an example, high-frequency subscriber sessions are responsible for only 0.55% of U_2, but their signaling overhead (ΔS) impacts are 8 times higher. For periodic sessions in U_2, the impacts are nearly 20 times higher.

To quantify disparity between the traffic volume and the resource consumption of high-frequency subscribers, we introduce the metric "payload density". Let C_{bytes} be the average payload size per session of a subscriber and $C_{signaling}$ be the total number of singling messages per session of the subscriber. We then compute the payload density, defined by $C_{bytes}/C_{signaling}$. Payload density is essentially one metric for measuring the effective data transfers per signaling message. Figure 3 plots the normalized payload density distributions of subscribers with different session activation frequency. We use the maximum payload density as the basis for normalizing. That is, normalized payload density of the subscriber i = *i's payload density/maximum payload density*. We observed that some active subscribers have an extremely high session-activate frequency, but a smaller payload density. The session activation frequency shows a negative correlation with the payload density.

From the operator's perspective, they charge only based on the traffic that subscribers have generated, and they prefer the situation of lower signaling cost but higher traffic volume; however, those active subscribers generate just little traffic and cause significant signaling load. From the perspective of other subscribers, in the process of generating the same amount of data traffic, those subscribers consume more resources, showing significant unfairness of resource consumption.

Table 1: Impact of high-frequency subscriber sessions.

Study scope	Contribution of high-frequency sessions			
	ΔV		ΔS	
	All	Periodicity	All	Periodicity
U_0 : *all subscribers*	100.00%	0.93%	-100.00%	-10.85%
$U_1 : N_{sa} \geq 5$	3.82%	0.55%	-19.46%	-8.61%
$U_2 : N_{sa} \geq 10$	1.66%	0.38%	-12.76%	-7.39%
$U_3 : N_{sa} \geq 15$	0.52%	0.25%	-8.28%	-5.96%
$U_4 : N_{sa} \geq 20$	0.30%	0.20%	-6.69%	-5.42%

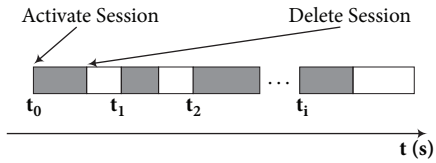

Figure 4: Model of subscriber session sequence.

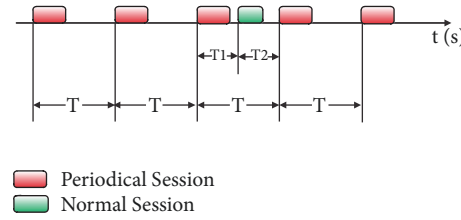

Figure 5: Model of SAI sequence.

Our estimation method has some limitations, because it neglects the signaling resources consumed in the session holding time. For example, during session holding time, RRC state transition may still occur (i.e., $DCH \longrightarrow FACH$), which brings a few round-trips of signaling messages. However, our method takes into account the signaling resources consumed in RRC connection setting up (i.e., $IDLE \longrightarrow DCH$) and the RRC connection release (i.e., $DCH \longrightarrow IDLE$ or $FACH \longrightarrow IDLE$). The recent cellular network measurement study [3] has demonstrated that the signaling messages of RRC connection setting up and release account for more than 60% of the total ones.

In summary, **some high-frequency users produce low data volume, but cause heavy signaling load, resulting in the unfairness of resource allocation and loss of network operators. Besides, with the increase of activation frequency, this phenomenon is more pronounced**.

5. Positive Correlation between Periodicity and Frequency

In this section, we focus on the characteristics of session activation about users described in previous section.

5.1. Detecting Periodicity in Session Activations. To detect the regularity of subscribers' session activation, we use DBSCAN algorithm [5] and a novel classification method to analyze subscribers' data sessions from a time perspective. Define high-frequency subscribers as $U = \{u_1, u_2, \ldots, u_n\}$, and each element of U represents a high-frequency subscriber. Here we use u_k as an example to illustrate our methods.

(i) Reconstruct all data sessions initiated by u_k in chronological order and model a sequence of sessions as shown in Figure 4.

(ii) Assuming $T_k = \{t_1, t_2, \ldots, t_n\}$ to represents the SAI sequence of u_k, we apply DBSCAN algorithm to cluster SAIs in T_k, and then map from the lowercase alphabet to these clusters in proper order. Thus, we get a cluster symbol series $C_k = c_1 \ldots c_m$, and c_i represents the symbol of the i^{th} cluster of u_k, i.e., $c_1 = a$ and $c_2 = b$.

(iii) Assume $TC_k = \{S(t_1), S(t_2), \ldots, S(t_n)\}$, in which the method S means the transition from t_i to the symbol of the cluster which t_i belong to. Then extract all subsequences of TC_k to find his self-patterns; i.e., if his subsequence is **"dbcb"**, then we retrieve all occurrences of **"dbcb"** in TC_k and calculate its ratio. If it exceeds the sequence fraction threshold (here, we set threshold to be 65%), then his self-pattern is **"dbcb"**; else consider his other subsequences.

(iv) To normalize self-activate pattern, we map from the capital alphabet to different cluster symbols of a particular pattern in a proper order. For example, suppose we have two self-patterns **"abca"** and **"dbcd"**, in this case, they are both mapped to **"ABCA"**, which indicates that they have the same normalized regularity.

In our method, we set sequence fraction threshold values for the consideration as follows: for each user, multiple applications can access network via cellular networks anytime, resulting in that a data session may contain multiple applications' data transfer, and each application may trigger a session activation if not existed. In this case, each application's access behavior, such as periodicity, may be hidden by others and make it harder to detect, as showed in Figure 5.

By making use of our periodicity detecting method, we tag two session types: **periodical sessions** which follow specific pattern, and **normal sessions** which behaviors are disorganized. And each user may have one or two session types. In our following sections, we sometimes analyze, respectively, with this definition.

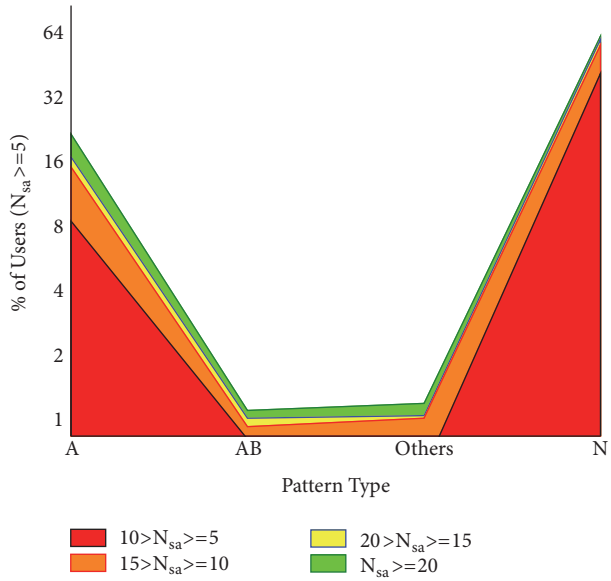

Figure 6: Distribution of subscribers over pattern type.

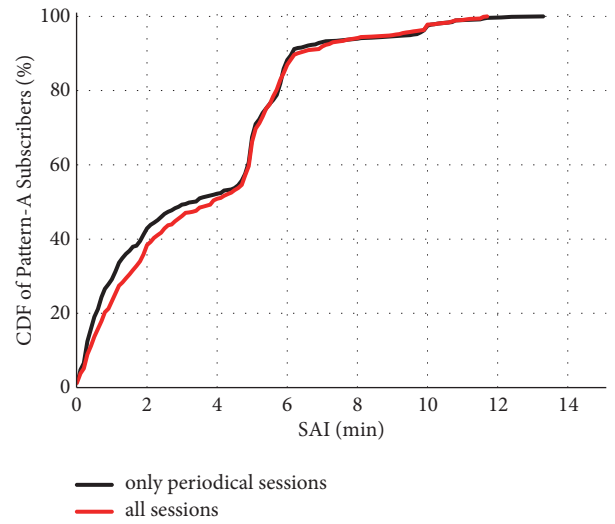

Figure 7: CDF of Pattern A subscribers over SAI.

5.2. The Phenomenon. Using the periodicity detecting method in previous section, we get tens session activation patterns. Here we choose some of them, which hold the largest proportion and illustrate them in Figure 6.

We use "Pattern X" to represent the regularity of session activations extracted by our periodicity detection method. Session activation behavior of a subscriber follow **Pattern A** means more than 50% of his SAIs are in accordance with one periodicity (T in Figure 5), and **Pattern AB** means more than 50% of the SAIs are in accordance with one or two periodicities (2T in Figure 5), as shown in (3). **"Others"** means there exist more than two periodicities in these sessions. Some other subscribers follow no pattern and tend to frequently activate sessions irregularly, which is represented as **Pattern N**.

$$t_i = \begin{cases} t_{i-1} + A, & if\ i\ is\ odd \\ t_{i-1} + B, & if\ i\ is\ even \end{cases} \quad (3)$$

We observed that **there is a positive correlation between the frequency of subscribers' session activation and the periodicity of SAI**. Higher frequency corresponds to a lower percentage of none-pattern traffic, which means the more frequently subscribers activate their sessions, the more possible that their activate sessions periodically. More than 40% of the subscribers ($N_{sa} \geq 10$) are in accordance with Pattern A, and for those with N_{sa} larger than 20, almost 70 % of them are Pattern A subscribers.

Figure 7 plots the CDF of Pattern A subscribers over SAI. The key observation is that several particular values dominate the intervals. We notice multiple small clusters, such as ≤ 3 minute, 4.5-6 minutes. Such values are likely to be set by mobile application developers in an ad hoc manner.

Comparing two curves, we find periodical sessions have a smaller SAI mean value.

5.3. Explanation (Application-Level Semantics). In order to find a reasonable explanation, we focus on the applications those subscribers applied, i.e., Facebook. To analyzing their application-level features, we respectively focus on Pattern A subscribers who trigger periodic session activations and Pattern N subscribers who generate nonperiodic sessions.

To carry out a deep investigation, we use not only the categories of network traffic characterized by port number but also application layer headers to distinguish the traffic from different applications [2]

5.3.1. Identifying Subscriber Behavior. Data flow, a specific data service between mobile and a fixed server, is typically identified by five tuples: src/dest address, src/dest ports, and protocol, for example, a tcp flow is represented by (∗,∗,∗,∗, TCP).

Each session contains multiple (≥ 0) flows, and flows within a session contain same local address (or protocol) and different server address. In this section, we identify each subscriber's application behavior as follows:

Assume each subscriber u has n sessions and is assigned a session vector $S_u = \{s_1, s_2, \cdots, s_n\}$, and each session s_i has m flows. The application type of m flows is stored in a vector $at(s_i) = (at_1, \ldots, at_m)$ ordered by flow index. Then u have many application type vectors and each vector may have duplicate entries. By merging vectors and removing duplicates, the p application types u used are stored in a vector $AT_u = \{at_1, at_2, \ldots, at_p\}$.

Here we define $\rho_{k,u}$ as the k_{th} application type fraction for user "u", then we can get

$$\rho_{u,k} = \sum_{i=1}^{n} \sum_{j=1}^{m} \begin{cases} \dfrac{1}{m \times n}, & if\ AT_{u,k} = at\left(s_i\right)_j \\ 0, & if\ AT_{u,k} \neq at\left(s_i\right)_j \end{cases} \quad (4)$$

When $\rho_{k,u}$ is much larger than others and AT_u have a small size, then the main application type of u is $AT_{u,k}$; i.e., if u has $AT_u = \{sns, im, game\}$ and $\rho_u = \{0.05, 0.02, 0.93\}$, then its type is "game". By the same token, larger AT_u

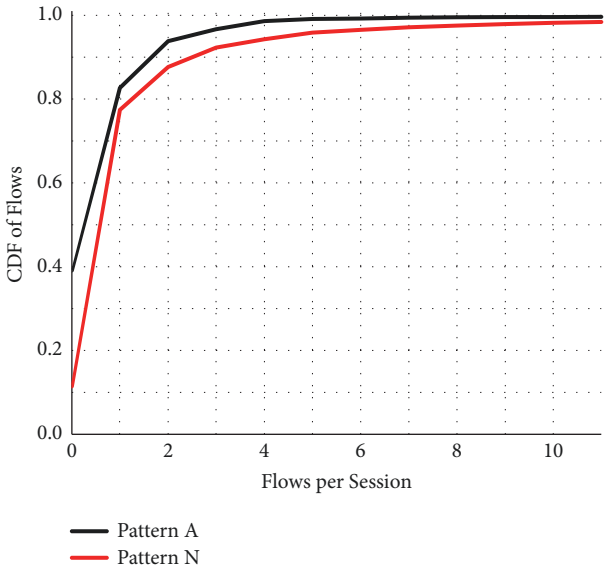

Figure 8: CDF of Sessions over the number of flows per session.

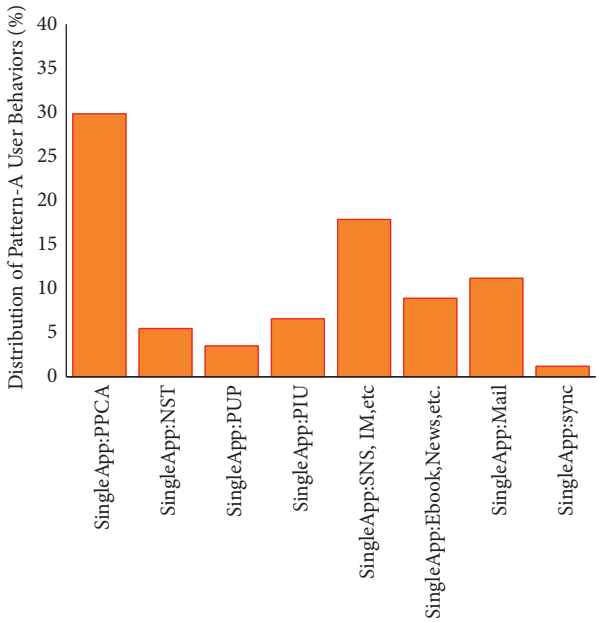

Figure 9: Distribution of Pattern A users over behavior type.

and lower ρ make subscribers a multiple-application type; i.e., if u has $AT_u = \{sns, im, game, news, sync\}$ and $\rho_u = \{0.2, 0.2, 0.2, 0.2, 0.2\}$, then its type is "multiple-application".

For reference, Table 2 lists important definitions used in this paper.

5.3.2. Application-Level Origin. As showed in Figure 8, we observed that most sessions have less than one flow, which means that to some extent a flow can represent a session. Besides, most periodical high-frequency users prefer to activate those zero- or 1-flow sessions. Then by calculating each user's application behavior with previous method, we identify more than 90% Pattern A users' behaviors, Figure 9 plots the distribution of behaviors by removing unknown types.

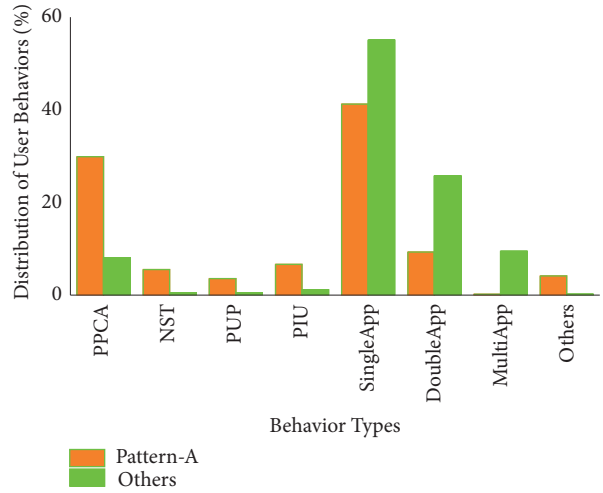

Figure 10: Distribution of high-frequency subscribers over behavior type.

We observed that most Pattern A users are SingleApp type, which means their features are mostly influenced by one specific application. By correlating multiple users' information with same application, we list features of some top behaviors as showed in Table 4.

6. Abnormal Behaviors

6.1. The Phenomenon. We find a certain correlation between the subscribers periodic session activations and abnormal behaviors (listed in Table 3). The overview of these apps or behaviors is illustrated in Figure 10. We observed that Pattern A subscribers (45.38% in sum) have a significant probability to trigger abnormal behaviors than others (9.95% in sum).

6.1.1. Periodical PDP Context Activation. Some subscribers periodically initiate PDP context with a fixed time interval, and then delete the PDP context soon. Figure 11 (red line) plots the CDF of Pattern A subscribers who perform this behavior over SAI. We observed that the SAI values is extremely fixed, one cluster range from 0min to 1 min, and another around 5min.

No data-plane transmission (no IP packets), short session duration, and fixed Session Activation Interval, obviously, it is enough to prove this behavior is suspicious. It would only cause the network to exchange signals continuously, wasting a lot of signaling resources without any actual utility (effective data transmission). In addition, the subscribers will not be charged with any fees based on data traffic accounting in this case.

6.1.2. Network-Side Automatic Termination. Some compromised mobile devices are about to access some premium or special websites continuously. As shown in Figure 12, we observe that these data flows have been successfully recognized by some middle boxes between the GGSN and

Table 2: Important definitions about user behavior.

Behavior Type	Definition
SingleApp	each user mainly applies one app type
DoubleApp	each user mainly applies two app types
MultiApp	each user applies more than two app types

Table 3: Typical abnormal behaviors.

Short Name	**Full Name**
PPCA	Periodical PDP Context Activation
NST	Network-side Automatic Termination
PUP	Periodical UDP Packets
PIU	Privacy Information Uploading

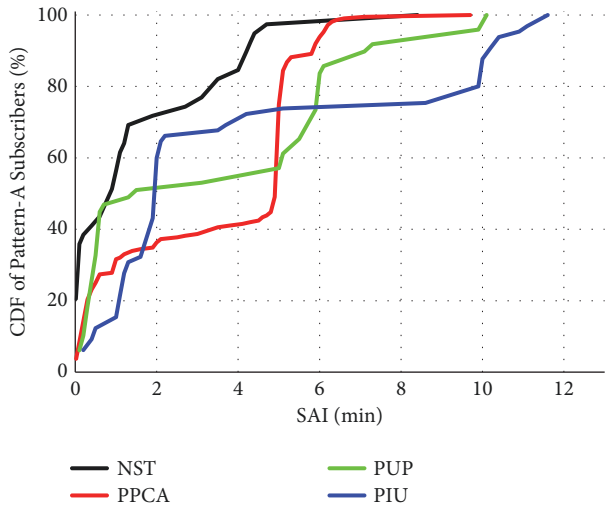

Figure 11: CDF of Pattern A subscribers over SAI.

the Internet such as Firewall and OCS (Online Charging System) where all traffic is composed of IP packets. Detected as abnormal access by firewall as lack necessary permissions, the connections will be deleted by the GGSN automatically. Although these methods prevent the target hosts or the core network from suffering invalid access, but the overconsumption of signaling resources cannot be avoided. Abnormal subscribers will continuously reactivate the PDP context and access as the automatic termination looks like a technical error for the subscribers, while GGSN will delete these PDP contexts immediately. Back and forth like this, a lot of signaling resources will be consumed.

Figure 11 (black line) plots the CDF of Pattern A subscribers who perform this behavior over SAI. We observed that major of this behavior may happen per (<1) min. Short period, high network resource consumption, and simple operation, make this behavior harmful to network and decrease others' experience.

6.1.3. Periodical UDP Packets. As showed in Figure 11 (green line), after creating the PDP sessions, the subscriber sends UDP packets to the target host in a certain period (from 1s to more than 30s or from 5min to 6min), and those sessions have its self-defined port, such as 8888, 12345. This type of sessions often has a relatively longer duration; however, the payload size and content of its packets mostly are fixed and fewer than dozens of bytes. Furthermore, we find some abnormal UDP broadcast behavior with few same content and high contracting frequency. As predicted in [6], periodical UDP packets may provide a way for the attackers to drain the battery power of subscribers' mobile devices by exploiting PDP context retention and the paging channel.

6.1.4. Privacy Information Uploading. The subscribers continuously initiate PDP contexts and upload privacy information [7], such as IMEI, phone numbers, etc. After finishing uploading, they disconnect themselves and after a while (SAI distribution plotted in Figure 11 (blue line)) repeat these operations. These behaviors may threaten the privacy security and at the same time waste signaling resources.

By identifying and clustering URLs, we model several typical requests, as showed in Table 5.

For those three model types, we observed major servers use the first type, such as "S2"; some UDP-based servers use the third model, only a few servers obtain users' privacy information by making use of the second model, such as a real request " http://S4:8080/?HOST=S2&R=/bs/&PhoneType=**&PhoneNumber=%2B861345&Version=1.6"

As we have studied in previous section, most sessions just have a flow and major Pattern A users are "SingleApp" type; in this section, we focus on flow level, for the consideration that a flow represent one-time data transfer between mobile and server.

Figures 13 and 14 plot the features of those servers (use "S1-22" to represent those servers) which get users' personal information.

We observed that there are dozens of servers obtain users' personal information, and major servers just obtain one or two privacy types. However, we still found that a few servers just request multiple times in each flow and do nothing, such as Server2, Server4, resulting in the sustained privacy loss.

7. Real-Life Implications of the Findings

It is necessary to reduce or manage the impact of high-frequency subscribers. Based on our study on the real trace

Table 4: Features of application behaviors.

Application Behavior	SAI Feature	Data Volume
SingleApp:PDP	fixed, multi-values (<1min or 5mins about)	zero
SingleApp:NST	fixed, small (<1.5mins)	very small
SingleApp:PUP	fixed, multi-values (<1min or 6mins about)	fixed, small
SingleApp:PIU	fixed, small (<2min)	fixed, small
SingleApp:IM,SNS	fixed, small (<1min)	disorganized
SingleApp:Mail	fixed, large (5mins about)	disorganized
SingleApp:Ebook, NEWS	disorganized	disorganized, large

Table 5: Typical privacy information obtain model.

URL Pattern
http://server/page?privacyKey1=privacyValue1&privacyKey2=privacyValue2
http://proxy/?HOST=server&R=relativePath&privacyKey1=privacyValue1&privacyKey2=privacyValue2http://proxy/?HOST=server&R =relativePath&privacyKey1=privacyValue1&privacyKey2=privacyValue2
http://server/purpose?name="privacyKey1" value="privacyValue1"

from an operational cellular network, we found that high-frequency subscribers can be extremely signaling resource-inefficient as they activate data sessions with high frequency but transfer few data in each session. As a result, it is necessary for cellular network operators to monitor the subscribers' session activation behaviors. We believe that this kind of solution will be critical for cellular service providers to improve the performance of resource allocation. Those subscribers with a high N_{sa} value and low payload density value should be paid close attention to.

8. Related Work

Servicing as the interface of network access, cellular network systems have experienced lots of versions to balance data-voice services and improve performance. However, the nature (or bottleneck) of these systems themselves has not changed, which means subscribers should consume network signaling resources for new network connections before access network, resulting in signaling overhead and even signaling storm [8].

In recent years, the characteristics of cellular data traffic and its impact on capacity planning, signaling cost and data transmission have attracted attention in the industrial circle. Previous studies can be classified approximately into several categories.

Network Architecture and Resource Management. RRC, which manages the handset radio interface, is the key coupling factor bridging the application traffic patterns and the lower-layer protocol behaviors [9]. Previous studies [10, 11] examine the RRC state machine and its interaction with cellular traffic for cellular networks. Also, some efforts [4, 12–17] measured various network performance metrics. Those studies investigated various aspects, such as crowed events, queue delay, and so on. Besides, Xu et al. characterized 3G data network infrastructure and found that the current routing of cellular data traffic was quite restricted [18]. Wang et al. unveiled cellular carriers NAT and firewall policies [19]. In this paper, by characterizing user behaviors, we took advantage of existing RRC state promotion, evaluated high-frequency access behavior's impact on signaling allocation/release, and found that invalid deployment and configuration of middle boxes will generate unwanted signaling traffic.

Mobile Behavior. The areas of traffic and application characteristic have recently received much attention by the research community, such as traffic dynamics [20, 21], geospatial dynamics [22], behavior patterns [23], mobility [24], and application usage patterns [25]. There are also several subscriber behavior studies based on deploying a custom logger on smartphones [26, 27]. Besides, by analyzing the periodicity of data transfer from ip-packet level, [4] studied periodic data transfer and its impact on resource consumption. In contrast, we analyzed periodicity of session activation from session level and perform the first multiangle investigation of high-frequency subscribers by rebuilding sessions, identifying session activation patterns and calculating application behavior for anonymized traces of an city-wide operational 3G network.

Unwanted Traffic, Detection, and Prevention. Complex and heavy signaling procedures render Internet-connected cellular networks vulnerable to a variety of abnormal data traffic [28–30]. Many studies have focused on various types of abnormal data, including virus [31, 32], spams [33], DoS attacks [34–36], phishing [37], charging [38, 39], etc. In response, significant work has been undertaken to detect [40, 41], model [42, 43], and defense [41] such problems. The goal of these studies is to design or model abnormal traffic to prevent legitimate use of data services [28, 30]. Unfortunately, few of these solutions have been widely deployed. In this paper, by collecting real-network data, modeling and extracting forensics, we do observe that some behaviors of high-frequency subscribers can lead to unwanted heavy signaling overloads.

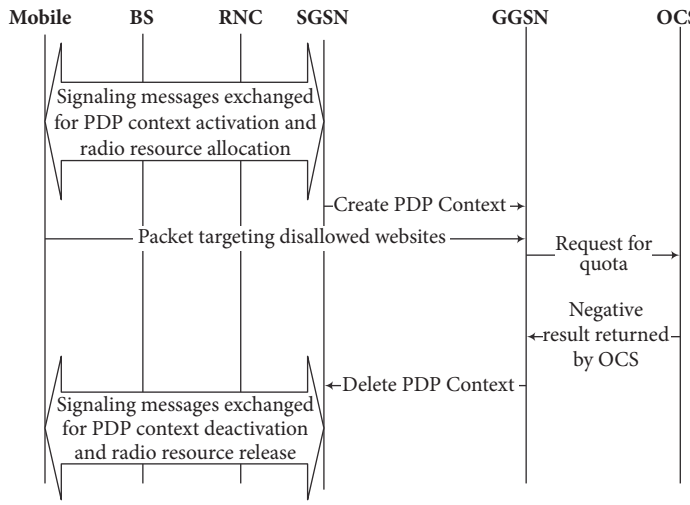

Figure 12: How OCS notifies GGSN to delete a PDP context.

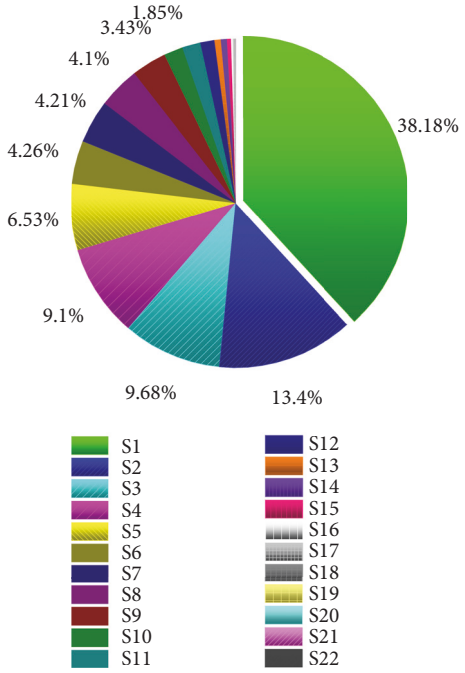

Figure 13: Distribution of Pattern A PIU users' flows over server.

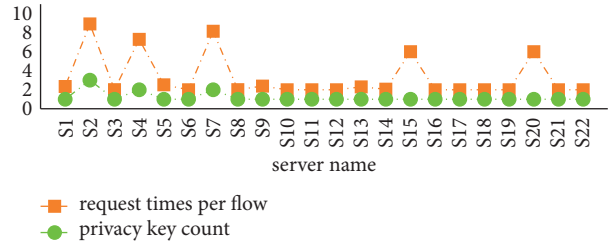

Figure 14: Analysis of Pattern A PIU users over request times per flow and personal key count.

Besides, prior efforts [4, 13, 44, 45] have explored that unwanted traffic can cause large-scale wastage of logical resources in cellular networks for various aspects, such as crowed events, periodic transfer, etc. In this paper, we proposed "high-frequency" traffic, a novel traffic type, verify, and enrich the above conclusion.

9. Conclusion

In this paper, we comprehensively characterized the impact and application-level origin of high-frequency subscribers in an operational cellular network in China. They consume much more signaling resources but have a lower utilization and arrive at super conclusion by making the following contributions:

(i) Inconsistent signaling-data bandwidth consumption, i.e., subscribers ($N_{sa} \geq 5$), generate 3.81% of the data traffic; however consume more than 19.46% of the total signaling resource, causing unfairness in charging.

(ii) Positive correlation between periodicity and frequency. Higher frequency corresponds to a lower percentage of none-pattern traffic.

(iii) Periodic subscribers tend to just apply one behavior or application, and amount of them actually does abnormal behaviors, such as periodical PDP context activation, network-side automatic termination, and privacy information uploading, and the payload density of these applications is extremely low.

10. Future Work

There are several directions for further research. First, we do not consider the impact of different kinds of applications. Second, how to reduce or manage the impact of high-frequency subscribers. We believe our findings in characterizing the session patterns of high-frequency subscribers directly have important implications on solutions to some of these issues.

Data Availability

The data comes from telecom operators, involving personal privacy of users. It can only be provided to specific people for academic research. Therefore, it cannot be provided.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work has been supported by National Natural Science Foundation of China (Grant no. 61671157). An earlier version of this paper was presented at "IFIP Networking Conference, 2013".

References

[1] R. P. Jover, L. P. Bloomberg, and N. York, "Some key challenges in securing 5G wireless networks," in *Proceedings of the FCC NOI DA*.

[2] J. Erman, A. Gerber, M. T. Hajiaghayi, D. Pei, and O. Spatscheck, "Network-aware forward caching," in *Proceedings of the 18th International World Wide Web Conference, WWW 2009*, pp. 291–300, Spain, April 2009.

[3] L. Qian, E. W. Chan, P. P. Lee, and C. He, "Characterization of 3G control-plane signaling overhead from a data-plane perspective," in *Proceedings of the Acm International Conference on Modeling*, p. 325, Paphos, Cyprus, October 2012.

[4] F. Qian, Z. Wang, Y. Gao et al., "Periodic transfers in mobile applications: Network-wide origin, impact, and optimization," in *Proceedings of the 21st Annual Conference on World Wide Web, WWW'12*, pp. 51–60, France, April 2012.

[5] M. Ester, H. P. Kriegel, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pp. 226–231, 1996.

[6] R. Racic, D. Ma, and H. Chen, "Exploiting MMS vulnerabilities to stealthily exhaust mobile phone's battery," in *Proceedings of the 2006 Securecomm and Workshops*, USA, September 2006.

[7] W. Enck, D. Octeau, p. McDaniel, and S. Chaudhuri, "A Study of Android Application Security , Usenix Conference on Security," in *Proceedings of the A Study of Android Application Security , Usenix Conference on Security*, vol. 2, 2011.

[8] "DoCoMo demands Google's help with signaling storm," http://www.rethink-wireless.com/2012/01/30/docomo-demands-googles-signalling-storm.htm.

[9] G. Gorbil, O. H. Abdelrahman, M. Pavloski, and E. Gelenbe, "Modeling and Analysis of RRC-Based Signalling Storms in 3G Networks," *IEEE Transactions on Emerging Topics in Computing*, vol. 4, no. 1, pp. 113–127, 2016.

[10] F. Qian, Z. Wang, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck, "Characterizing radio resource allocation for 3G networks," in *Proceedings of the ACM SIGCOMM 10th Internet Measurement Conference (IMC '10)*, pp. 137–150, ACM, November 2010.

[11] J. Huang, F. Qian, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck, "A close examination of performance and power characteristics of 4G LTE networks," in *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services (MobiSys '12)*, pp. 225–238, ACM, June 2012.

[12] X. Liu, A. Sridharan, S. Machiraju, M. Seshadri, and H. Zang, "Experiences in a 3G Network: Interplay between the Wireless Channel and Applications," in *Proceedings of the ," Acm International Conference on Mobile Computing and Networking*, vol. 239, no. 7, pp. 211–222, September 2008.

[13] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, S. Venkataraman, and J. Wang, "A first look at cellular network performance during crowded events," in *Proceedings of the 2013 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, SIGMETRICS 2013*, pp. 17–28, June 2013.

[14] J. Huang, F. Qian, Y. Guo et al., "An in-depth study of LTE," *Computer Communication Review*, vol. 43, no. 4, pp. 363–374, 2013.

[15] P. Benko, G. Malicsko, and A. Veres, "A large-scale, passive analysis of end-to-end TCP performance over GPRS," in *Proceedings of the IEEE INFOCOM 2004 - Conference on Computer Communications - Twenty-Third Annual Joint Conference of the IEEE Computer and Communications Societies*, pp. 1882–1892, China, March 2004.

[16] H. Jiang, Y. Wang, K. Lee, and I. Rhee, "Tackling bufferbloat in 3G/4G networks," in *Proceedings of the ACM Internet Measurement Conference (IMC '12)*, pp. 329–342, November 2012.

[17] O. H. Abdelrahman, "Detecting network-unfriendly mobiles with the random neural network," *Probability in the Engineering and Informational Sciences*, vol. 30, no. 3, pp. 514–531, 2016.

[18] Q. Xu, J. Huang, Z. Wang, F. Qian, A. Gerber, and Z. M. Mao, "Cellular data network infrastructure characterization and implication on mobile content placement," *ACM SIGMETRICS Performance Evaluation Review*, vol. 39, no. 1, p. 277, 2011.

[19] Z. Wang, Z. Qian, Q. Xu, Z. Mao, and M. Zhang, "An untold story of middleboxes in cellular networks," *Computer Communication Review*, vol. 41, no. 4, p. 374, 2011.

[20] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci, "Measuring serendipity: connecting people, locations and interests in a mobile 3G network," in *Proceedings of the 2009 9th ACM SIGCOMM Internet Measurement Conference, IMC 2009*, pp. 267–279, USA, November 2009.

[21] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang, "Characterizing and modeling internet traffic dynamics of cellular devices," in *Proceedings of the Acm Sigmetrics Joint International Conference on Measurement and Modeling of Computer Systems*, vol. 39, no.1, pp. 305–316, San Jose, California, USA, June 2011.

[22] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Characterizing geospatial dynamics of application usage in a 3G cellular data network," in *Proceedings of the IEEE Conference on Computer Communications, INFOCOM 2012*, pp. 1341–1349, USA, March 2012.

[23] R. Keralapura, A. Nucci, Z. Zhang, and L. Gao, "Profiling users in a 3g network using hourglass co-clustering," in *Proceedings of the International Conference on Mobile Computing and Networking*, p. 341, Chicago, Illinois, USA, September 2010.

[24] E. Halepovic and C. Williamson, "Characterizing and modeling user mobility in a cellular data network," in *Proceedings of the PE-WASUN'05 - Second ACM International Workshop on Performance Evaluation of Wireless Ad Hoc, Sensor, and Ubiquitous Networks*, pp. 71–78, Canada, October 2005.

[25] Q. Xu, J. Erman, A. Gerber, Z. Mao, J. Pang, and S. Venkataraman, "Identifying diverse usage behaviors of smartphone apps," in *Proceedings of the ACM SIGCOMM Internet Measurement Conference (IMC '11)*, pp. 329–344, November 2011.

[26] H. Falaki, D. Lymberopoulos, R. Mahajan, S. Kandula, and D. Estrin, "A first look at traffic on smartphones," in *Proceedings*

of the 10th Internet Measurement Conference (IMC '10), pp. 281–287, Melbourne, Australia, November 2010.

[27] H. Falaki, R. Mahajan, S. Kandula, D. Lymberopoulos, R. Govindan, and D. Estrin, "Diversity in smartphone usage," in *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services (MobiSys '10)*, pp. 179–194, ACM, San Francisco, Calif, USA, June 2010.

[28] P. Traynor, P. McDaniel, and TL. Porta, "On Attack Causality in Internet-connected Cellular Networks," in *Proceedings of the USENIX Security Symposium*, 2007.

[29] J. Serror, Z. Hui, and J. C. Bolot, "Impact of paging channel overloads or attacks on a cellular network," in *Proceedings of the WiSE 2006 - 5th ACM Workshop on Wireless Security*, pp. 75–84, USA, September 2006.

[30] P. P. Lee, T. Bu, and T. Woo, "On the Detection of Signaling DoS Attacks on 3G Wireless Networks," in *Proceedings of the IEEE INFOCOM 2007 - 26th IEEE International Conference on Computer Communications*, pp. 1289–1297, Anchorage, AK, USA, May 2007.

[31] N. Leavitt, "Mobile phones: the next frontier for hackers?" *The Computer Journal*, vol. 38, no. 4, pp. 20–23, 2005.

[32] J. Cheng, S. H. Y. Wong, H. Yang, and S. Lu, "SmartSiren: virus detection and alert for smartphones," in *Proceedings of the 5th International Conference on Mobile Systems, Applications and Services (MobiSys '07)*, pp. 258–271, ACM, San Juan, Puerto Rico, June 2007.

[33] A. P. Felt, M. Finifter, E. Chin, S. Hanna, and D. Wagner, "A survey of mobile malware in the wild," in *Proceedings of the 1st ACM Workshop on Security and Privacy in Smartphones and Mobile Devices (SPSM '11) held in Association with the 18th ACM Conference on Computer and Communications Security (CCS '11)*, pp. 3–14, October 2011.

[34] P. Traynor, M. Lin, M. Ongtang et al., "On cellular botnets: Measuring the impact of malicious devices on a cellular network core," in *Proceedings of the 16th ACM Conference on Computer and Communications Security, CCS'09*, pp. 223–234, USA, November 2009.

[35] W. Enck, P. Traynor, P. McDaniel, and T. La Porta, "Exploiting open functionality in SMS-capable cellular networks," in *Proceedings of the 12th ACM Conference on Computer and Communications Security (CCS '05)*, pp. 393–404, November 2005.

[36] B. Zhao, C. Chi, W. Gao, S. Zhu, and G. Cao, "A chain reaction DoS attack on 3G networks: Analysis and defenses," in *Proceedings of the 28th Conference on Computer Communications, IEEE INFOCOM 2009*, pp. 2455–2463, Brazil, April 2009.

[37] W. Dan, *Smartphone security: Trends and predictions, Secure Application Development*, 2011.

[38] C. Peng, C. Li, G. Tu, S. Lu, and L. Zhang, "Mobile data charging," in *Proceedings of the ACM conference on Computer and communication security*, p. 195, Raleigh, North Carolina, USA, October 2012.

[39] C. Peng, G. Tu, C. Li, and S. Lu, "Can we pay for what we get in 3G data access?" in *Proceedings of the International Conference on Mobile Computing and Networking*, p. 113, Istanbul, Turkey, August 2012.

[40] O. H. Abdelrahman, E. Gelenbe, G. Görbil, and B. Oklander, "Mobile Network Anomaly Detection and Mitigation: The NEMESYS Approach," in *Information Sciences and Systems 2013*, vol. 264 of *Lecture Notes in Electrical Engineering*, pp. 429–438, Springer International Publishing, Cham, 2013.

[41] R. Kaur, M. Gaur, L. Suresh, and V. Laxmi, "DoS Attacks in MANETs," in *Cyber Security, Cyber Crime and Cyber Forensics*, Advances in Digital Crime, Forensics, and Cyber Terrorism, pp. 124–145, IGI Global, 2011.

[42] G. Maciá-Fernández, J. E. Díaz-Verdejo, and P. García-Teodoro, "Mathematical model for low-rate dos attacks against application servers," *IEEE Transactions on Information Forensics and Security*, vol. 4, no. 3, pp. 519–529, 2009.

[43] Y. Tang, X. Luo, Q. Hui, and R. K. C. Chang, "Modeling the vulnerability of feedback-control based internet services to low-rate DoS attacks," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 3, pp. 339–353, 2014.

[44] F. Ricciato, P. Svoboda, E. Hasenleithner, and W. Fleischer, "On the impact of unwanted traffic onto a 3G network," in *Proceedings of the 2nd International Workshop on Security, Privacy and Trust in Pervasive and Ubiquitous Computing, SecPerU 2006*, pp. 49–56, France, June 2006.

[45] F. Qian, Z. Wang, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck, "TOP: Tail optimization protocol for cellular radio resource allocation," in *Proceedings of the 18th IEEE International Conference on Network Protocols, ICNP'10*, pp. 285–294, Japan, October 2010.