Soft Computing (2024) 28:1865–1881
https://doi.org/10.1007/s00500-023-09486-x

FOUNDATION, ALGEBRAIC, AND ANALYTICAL METHODS IN SOFT COMPUTING

A formal security analysis of the fast authentication procedure based on the security context in 5G networks

Zhiwei Cui[1] · Baojiang Cui[1] · Li Su[2] · Haitao Du[2] · Jie Xu[1] · Junsong Fu[1]

Accepted: 16 November 2023 / Published online: 10 January 2024

Abstract

The security context, generally stored in the universal subscriber identity module card or the baseband chip, is the critical information applied by the subscriber to access the 5G network during the fast authentication procedure. Once exposed or illegally used, the security context can be exploited to derive various keys for authentication and encryption. Despite its importance, challenges and questions still remain in the previous relevant research. To fill this gap, by adopting the security protocol verification tool ProVerif, we provide a comprehensive formal model of the fast authentication procedure based on the security context to analyze whether security goals can be met. Unfortunately, we uncover two vulnerabilities, including one never reported before. Our analysis shows that these vulnerabilities stem from fundamental design flaws in the cellular network protocol and thus apply to the 4G network. These vulnerabilities could be exploited to launch several attacks, including impersonation and eavesdropping. We have validated these attacks using 5 mobile phones from 5 different baseband manufacturers through experimentation in three mobile carriers. We find an insecure implementation of one of these phones, which exposed it to replay attacks. And we further discuss the security threats posed by the impersonation attack, such as location spoofing and one-tap authentication bypass, which is verified on 10 popular apps. We finally propose several countermeasures to eliminate these security issues. Actually, we have reported the novel vulnerability to the GSM Association and received a confirmation in the form of a coordinated vulnerability disclosure (CVD) number CVD-2022-0057.

Keywords Security context · 5G network · Vulnerabilities · Attacks

1 Introduction

5G is the state-of-the-art cellular network with the promise of ultra-low latency and enables plenty of emerging applications, such as autonomous driving, smart healthcare, and industrial IoT. According to statistics, 5G networks have 1.5 billion subscribers by 2022 (GSMA 2022). They access the 5G network via their universal subscriber identity module (USIM) cards to consume services, such as voice and short message service (SMS). Subscribers are protected by many security mechanisms, the most important of which is the Authentication and Key Agreement (AKA) procedure. The procedure utilizes the pre-key shared by the USIM card and the network to enable them to mutually authenticate each other and establish the session keys to secure communications.

Given its importance, the AKA procedure has inspired a lot of enthusiasm among researchers. Basin et al. (2018) provided the first comprehensive formal model of 5G AKA. The

✉ Baojiang Cui
cuibj@bupt.edu.cn

Zhiwei Cui
zwcui@bupt.edu.cn

Li Su
suli@chinamobile.com

Haitao Du
duhaitao@chinamobile.com

Jie Xu
cheer1107@bupt.edu.cn

Junsong Fu
fujs@bupt.edu.cn

[1] Beijing University of Posts and Telecommunications, Beijing 100876, China

[2] China Mobile Research Institute, Beijing 100876, China

logical vulnerabilities in the AKA procedure can be exploited to launch several attacks, including identification (Shaik et al. 2019), activity monitoring (Borgaonkar et al. 2019), location tracking (Shaik et al. 2016), etc. To protect the privacy of subscribers, several security enhancements to the AKA protocol are proposed (3GPP 2018; Wang et al. 2021). However, a perfect AKA procedure may not ensure a secure access to the network for the subscriber.

A key observation is that a mobile equipment (ME) equipped with a valid USIM card can connect to the network without going through an AKA procedure. Specifically, the mobile phone does need to process an AKA procedure when it launches the initial registration procedure. Usually, a security context is created as the result of the AKA procedure and maintained by the network and the user equipment (UE) consisting of the USIM card and the ME (3GPP 2022c). When the UE reconnects to the network, the UE sends an encrypted and integrity-protected Registration Request message based on the cached context. If the network finds a corresponding security context, it can bypass the AKA procedure (3GPP 2022c). In our work, we name it the fast authentication procedure. Once the security context can be illegally used, an attacker may connect to the network as a victim. It is hence necessary to investigate the existing design and implementation of the security context in the 5G network.

For one UE, the security context can be stored in the USIM card or the ME. If the USIM does not support storing the security context, the baseband chip in the ME stores the context together with the permanent identity. The 3GPP group standards several security mechanisms to defend against attacks on the security context, including PIN-based access control (3GPP 2022b) and identity verification (3GPP 2022e). Whether these mechanisms can achieve the intended purpose and whether there will be shortcomings in the implementation stage are still worth discussing.

Unfortunately, recent research yields a negative answer (Zhao et al. 2021). The PIN-based access control may expose the security context stored in the USIM card to an attacker. Once exposed, the attacker can launch some attacks, such as traffic eavesdropping and impersonation. This proves that existing security mechanisms are flawed and cannot provide effective protection to defend against attacks on the security context. However, the research is still of some limitations: (1) ignore the situation that the security context is stored in the baseband chip; (2) the vulnerability was discovered through the manual analysis.

In our work, we apply the formal method to determine whether security goals are met by these security mechanisms proposed by the 3GPP group. The method has been used in assessing the security of protocols, such as 5G authentication (Basin et al. 2018), TLS (Beurdouche et al. 2015), etc. To provide a more realistic model, we first extract security assumptions and goals of the fast authentication procedure from the specifications and the mobile carriers' practices. Then, we provide a comprehensive formal model of this procedure using ProVerif (Blanchet et al. 2018). We propose some solutions to address the challenges in the modeling procedure. For example, we use the map data structure to indicate that a valid PIN is required to read the security context in the USIM card. The formal analysis results show that there are vulnerabilities in the protection mechanism of the security context in both storage scenarios. Further analysis shows that these protocol vulnerabilities also exist in 4G networks. Specifically, when the security context exists in the USIM card, the attacker can read the context through hardware or software (Zhao et al. 2021). In our work, we exploit this vulnerability to propose two more effective attacks (i.e., impersonation and eavesdropping), which need one-time physical access. Then, we point out that these attacks using the security context stored in the USIM card can be effective for a long time. When the security context is stored in the baseband chip, the identity verification can be bypassed. The novel vulnerability could lead to the impersonation attack. To ensure that these attacks are indeed realizable in practice, we have responsibly confirmed them through experimentation using 5 mobile phones from 5 different baseband manufacturers in three commercial mobile carriers. In the experiment, we also find some insecure implementations of handling the security context on one of these phones and several open-source 4G and 5G networks. These unsafe practices could lead to the Denial of Service (DoS) attack and the replay attack. Then, we further discuss security threats caused by the impersonation attack, including location spoofing and one-tap authentication (OTAuth) bypass leading to privacy disclosure. According to statistics, more than 400 apps support the OTAuth service. We have performed relevant verification on 10 popular apps, such as Alipay, TikTok, and Weibo. Finally, we offer several recommendations for the identified security issues from four perspectives.

In summary, our contributions are as follows:

1. We provide a comprehensive formal model for the fast authentication procedure using the security context in 5G networks. And we identify two vulnerabilities, one of which is novel.
2. We present two attacks exploiting the identified vulnerabilities and verify them in three carriers. In addition, we analyze the serious security threats posed by the impersonation attack.
3. We discover many incorrect handlings of the security context on a commercial mobile phone and some open-source projects.
4. We analyze root causes and propose several effective countermeasures to prevent these attacks.

The rest of this paper is organized as follows. We present the preliminaries of the USIM card, the 5G registration procedure, and the 5G security context that are relevant to our work in Sect. 2. Section 3 introduces the formal model of the fast authentication procedure based on the security context in 5G networks. We describe the analysis results and the identified vulnerabilities, and evaluate the impact of these vulnerabilities on 4G networks in Sect. 4. Then we design several attacks as described in Sect. 5. Section 6 further discusses security threats caused by the impersonation. Several countermeasures are presented in Sect. 7. We summarize the related work in Sect. 8. At last, Sect. 9 concludes this paper.

Ethics discussion We perform all the experiments in a responsible way. All USIM cards and mobile phones involved in the experiments are our own to prevent affecting other users. And we put the simulated network in a radio isolated shield box to prevent the exposure of signals. We have notified mobile carriers about the unsafe implementations. In addition, we have submitted the novel vulnerability to the GSM Association (GSMA) for mitigation and obtained a coordinated vulnerability disclosure (CVD) number CVD-2022-0057, details of which are not publicly available (GSMA 2023). At the same time, we have reported these vulnerabilities in handling the security context to the China National Vulnerability Database (CNVD) and proposed mitigation measures. Unisoc promises to respond a Common Vulnerabilities and Exposures (CVE) and publicly acknowledge our work on its official website (Unisoc 2023).

This paper is an extended version of our previous work (Cui et al. 2023) accepted at the 6th International Symposium on Mobile Internet Security (MobiSec 2022). The main differences between these two versions are listed below: (1) we distinguish the security context under the two storage scenarios and refine our formal model. The analysis results show that the victim faced different attacks in these two scenarios. (2) We present a novel eavesdropping attack, which can be exploited to further launch the Domain Name System (DNS) redirection attack. (3) Detailed analysis shows that the discovered vulnerabilities and the proposed attacks are applicable to 4G networks. (4) We have conducted relevant tests on 5 mobile phones equipped with different baseband manufacturers. And we identify several security flaws in one of the phones and many open-source projects. (5) We describe the principle of the OTAuth service. Exploiting the impersonation attack, an attacker can log in to hundreds of app accounts of the victim. We have tested 10 popular apps that support the OTAuth service. (6) We discuss the persistence of attacks based on the security context stored in the USIM card. It turns out that these attacks requiring the one-time physical access can last for a long time. These attacks can still be carried out even if the attacker has been away from the victim.

2 Background

2.1 USIM

Each USIM card has a unique integrate circuit card identity (ICCID). The USIM card contains various critical information for a subscriber to access 4G/5G networks, including the identity, the permanent key, the phone number, etc. These parameters are stored in the form of files. Each file has a unique ID and can be set with different access conditions managed by the mobile carrier. The 3GPP group defines four access modes (3GPP 2022b; Zhao et al. 2021): (1) always (ALW), with no restrictions; (2) never (NEV), inaccessible; (3) ADM, where the correct administrative key owned by the mobile carrier needs to be verified; (4) PIN, where the correct PIN needs to be entered unless the verification is disabled. The number of consecutive incorrect PIN entries is limited. If the limit is exceeded, the USIM card will be automatically locked or destroyed. The read and update conditions of one file may be different. For example, the read condition of the International Mobile Subscriber Identity (IMSI) with ID 6F07 is the PIN mode, but its update condition is set to the ADM mode.

The USIM card provides the standard application programming interface (API). A mobile phone can send the application protocol data unit (APDU) to the USIM card to read or update files through the interface. In addition, the USIM card supports some security functions (e.g., authentication and key derivation), which can be used by sending APDUs.

2.2 5G registration procedure

As shown in Fig. 1, the 5G registration procedure involves four entities, namely the USIM card, the ME, the base station (BS), and the access and mobility management function (AMF). The BS provides the wireless network for the ME and the AMF to communicate with each other. The AMF is responsible for the mutual authentication with subscribers. This procedure includes the initial registration procedure and the fast authentication procedure. We briefly introduce the above two procedures, respectively, in the following.

Initial registration procedure For privacy-preserving, the ME sends an initial Registration Request message containing the Subscription Concealed Identifier (SUCI) instead of the Subscription Permanent Identifier (SUPI) (3GPP 2022e). After checking if the identity is valid, the AMF will initiate the AKA procedure. Specifically, the AMF generates an Authentication Request message and sends it to the ME. The message contains a random number (RAND) and an authentication token (AUTN) generated by several secure functions that take the permanent key and the RAND as parameters. Once receiving the Authentication Request message, the ME

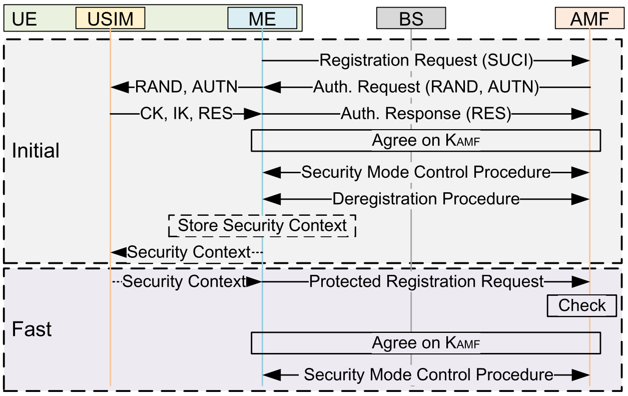


Fig. 1 5G registration procedure

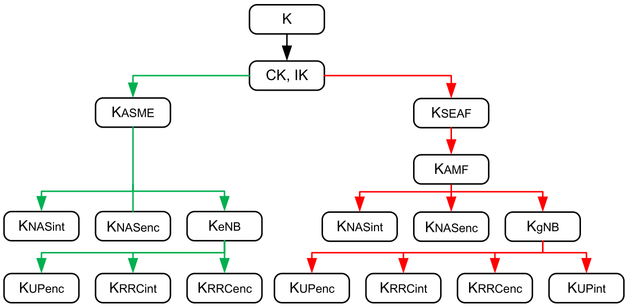


Fig. 2 Key derivation. The red line (resp. the green line) represents key derivation for 5G (resp. 4G) (color figure online)

forwards these two parameters to the USIM card in the form of APDU. The USIM card calculates the cipher key (CK), the integrity key (IK), and the authentication response (RES) through the permanent key and security algorithms preset in the USIM card, and returns them to the ME. Then the ME and the AMF agree on the K_{AMF}, which can be used to derive several keys for data encryption and integrity protection as shown in Fig. 2. Next, the two entities negotiate encryption and integrity protection algorithms through the security mode control procedure. Finally, when the ME initiates the deregistration procedure (e.g., turn on the airplane mode and power off), a security context is maintained by the UE and the AMF simultaneously. For the UE, the security context is stored in the ME or the USIM card.

Fast authentication procedure To optimize the re-registration procedure, the 3GPP group standards the fast authentication procedure based on the security context (3GPP 2022c). If the UE has a security context, it shall send a protected Registration Request message with the 5G-Globally Unique Temporary Identifier (5G-GUTI) as identity, not the SUCI. The complete message shall be ciphered in a Non-Access-Stratum (NAS) container and shall also be integrity protected with the security context (3GPP 2022c). In the case that the AMF has the same security context and successfully decrypts the message, which passes the integrity check, the AKA procedure may be omitted. Then the UE and the AMF derive the keys required for subsequent communication based on this security context to activate security. It's worth noting that the 3GPP group does not mandate skipping the AKA procedure, which depends on the mobile carriers. The three carriers involved in our work all support the fast authentication procedure.

2.3 5G security context

The 5G security context includes the 5G-GUTI and the NAS security context. If the USIM supports storing the 5G security context, the 5G-GUTI, and the NAS security context are stored in file EF5GS3GPPLOCI (ID: 4F01) and file EF5GS3GPPNSC (ID: 4F03) respectively. The read and update conditions for the two files are set to the PIN mode. Otherwise, they shall be stored in the non-volatile memory of the baseband chip together with the permanent identity of the subscriber.

The NAS security context contains the K_{AMF} with the associated key set identifier ngKSI, the UE security capabilities, and the uplink and downlink NAS COUNT values (3GPP 2022b). The ngKSI has 8 values from 0 to 7. When its value is 7, it means that the UE has no security context. The NAS COUNT value is used to resist replay attacks and independently maintained by the UE and the AMF. The NAS COUNT value is constructed as an 8-bit NAS sequence number (SQN) and a 16-bits NAS overflow counter. Their initial values are 0. After sending each new protected NAS message, the NAS SQN shall be increased by one, and if the result is zero (due to wrap around), the NAS overflow counter shall also be incremented by one. The UE security capabilities identify the selected NAS integrity and encryption algorithms. Using these algorithms, the UE derives encryption and integrity protection keys from the K_{AMF} and the NAS COUNT values to protect the Registration Request message. The message contains the 5G-GUTI and the ngKSI in plaintext, which uniquely identify a security context. The AMF queries the same security context through the two values.

To prevent the security context from being used illegally, the 3GPP group has proposed some security mechanisms: (1) the PIN-based access control prevents malicious reading and updating of the security context stored in the USIM card (3GPP 2022b); (2) to use the security context in the baseband chip, the permanent identity stored in the USIM card inserted into the ME shall match the permanent identity corresponding to the security context (3GPP 2022e).

3 Formal modeling the fast authentication procedure in 5G networks

In this section, we introduce the formal model of the fast authentication procedure based on the security context in 5G networks. First, we derive precise security assumptions and goals. Then, we introduce the modeling of security goals. Finally, we present the challenges and corresponding solutions in modeling the fast authentication procedure.

3.1 More precise security assumptions

The 3GPP specifications define several security assumptions (3GPP 2022b, c, e). However, the assumptions are very strong, and many deployments of mobile carriers do not strictly follow them. To provide a more precise and comprehensive analysis, we combine the information given in the 3GPP documents and real-world scenarios to define security assumptions. In addition, since automatically extracting security assumptions from ambiguous and complex documents is very challenging, we manually perform this work.

Assumptions on channels The public wireless channel between the UE and the AMF faces security threats posed by two types of attackers. An active attacker can set up malicious entities (e.g., to impersonate the AMF or the UE) to send or receive messages. A passive attacker can eavesdrop on the communications on the channel, but cannot decrypt the encrypted messages and never send a message. In our work, we consider active attackers. From now on, the attacker refers to an active attacker unless explicitly stated otherwise.

Assumptions on cryptographic primitives The fast authentication procedure involves multiple cryptographic algorithms, including confidentiality protection, integrity protection, and key derivation functions. We assume that all cryptographic algorithms are public and secure. In our analysis, we consider that these functions are adopted and have secure implementations. An attacker without the correct key cannot encrypt or decrypt the message and pass the integrity check.

Assumptions on entities In our research, we consider two entities, namely the UE and the AMF. In the fast authentication procedure, the UE and the AMF perform mutual authentication through the security context. If the authentication is successful, the AMF will send the K_{gNB} to the BS to protect the communication between the UE and the BS. However, before this, the BS just acts as a bridge for the UE to communicate with the AMF. Therefore, we omit the modeling of the BS without affecting the security analysis results for the fast authentication procedure to simplify the work. We assume that the UE and the AMF jointly maintain a valid security context. The UE consists of the USIM card and the ME as mentioned above. And the attacker has physical access to the USIM and the ME. In real life, there are indeed chances for other people to physically access the UE. For example, when a confidential meeting is held, the relevant personnel are required to put their mobile phones together outside the meeting room.

Assumptions on data protection We assume the following data are public and available to the attacker: the default PIN and the SUPI. The three mobile carriers involved in our work set the default PIN to 1234 and disable the SUPI protection mechanism. The SUPI type is set to the IMSI. As a result, the SUPI will be transmitted in plaintext and obtained by the attacker (Nie et al. 2022).

3.2 Security goals in the fast authentication procedure

We now describe the security goals that the fast authentication procedure based on the security context should achieve.

Authentication properties The 3GPP group makes claims about the 5G subscriber authentication properties in the document. We have identified the relevant statements and translated them into formal security goals. These goals can be precisely formalized using Lowe's taxonomy of authentication properties, which specifies four levels of authentication between two agents A and B from A's point of view (Lowe 1997): (1) aliveness, which ensures that B has run the protocol previously, but not necessarily with A; (2) weak agreement, which ensures that B and A have run the protocol previously, but not necessarily with the same data. This prevents impersonation attacks; (3) non-injective agreement, which ensures that B and A have run the protocol previously and agree on the data. This prevents message tampering attacks; (4) injective agreement, which ensures that for each run of the protocol of A, there is a unique matching run of B on the basis of non-injective agreement. This prevents replay attacks. The UE and the AMF shall be able to mutually authenticate, which is ensured by the key confirmation based on the security context. And the replay protection shall be supported for the NAS signaling messages both in the AMF and the UE through the NAS COUNT values. Formally, the UE must obtain an injective agreement on the fast authentication procedure with the UE. Conversely, so does the AMF.

Confidentiality properties While it is not specified, obviously the confidentiality of the security context is required. The 3GPP group proposes several protection measures for the security context stored in the USIM card or the baseband chip (3GPP 2022b, e). Since all the data in a security context except the K_{AMF} and the 16-bit NAS overflow counter are transmitted in plaintext or are deterministic values (e.g., the security algorithms selected by the mobile carrier are generally unchanged), the two data should be guaranteed to be confident. Formally, the cryptographic key K_{AMF} and the counter protected by these measures should remain secret in the presence of the attacker.

Privacy properties The fast authentication procedure shall ensure that the privacy associated with a subscriber cannot be compromised so that an attacker cannot identify or track the subscriber. The 5G-GUTI that identifies the subscriber identity will be updated in each registration procedure. Therefore, an attacker cannot identify the subscriber by the 5G-GUTI. However, attackers can obtain the unprotected SUPI through user location attacks similar to IMSI-catcher attacks (Shaik et al. 2016). Since the location privacy leakage is a well-known problem, we disregard the privacy property in our work.

3.3 Modeling the fast authentication procedure in ProVerif

To analyze the fast authentication procedure, we used the ProVerif prover (Blanchet et al. 2018), which is a protocol verification tool to prove the security properties of protocols, including authentication properties, confidentiality properties, and privacy properties. It has been used in verifying real-world protocols (Basin et al. 2018; Beurdouche et al. 2015). It is capable of handling many cryptographic algorithms, such as the public-key cryptography and the hash function. Although ProVerif does not support XOR operation, which is important for 5G key derivation, it can still be used to verify the fast authentication procedure. Since we assume that the cryptographic algorithm is secure, we can simplify the key derivation procedure.

```
type key.
fun senc(bitstring,key):bitstring.
reduc forall m:bitstring,k:key;
 sdec(senc(m,k),k)=m.
```

ProVerif can automatically deduce the logical derivation of the security goals to be proved based on the formal description of the protocol. If a security goal is violated, it can give a detailed attack. In ProVerif, messages are described as terms. A term is constructed by constructors. Taking symmetric encryption as an example, we define the type key, and *senc(m, k)* represents the message m encrypted using the *k*, where *senc(bitstring, key)* is a constructor. And the equation *sdec(senc(m, k), k) = m* represents the decryption of the plaintext with the same *k*.

3.3.1 Formalizing security goals

ProVerif can prove reachability properties, correspondence assertions, and observational equivalence.

Authentication properties can be checked via the correspondence assertions. The correspondence assertions are used to capture relationships between events. It can be expressed that if a specified event with some arguments has been executed, then the other with the same arguments event has been previously executed. For example, if the AMF is able to verify an encrypted and integrity-protected Registration Request message with information elements (including 5G-GUTI, ngKSI, uplink NAS SQN), NAS container, and message authentication code (MAC) (event *amf_verify*), then it means that the UE has initiated a unique Registration Request message with the same parameters (event *ue_init*). We can use the following query to check the injective agreement on these data.

```
query x:ies,y:container,z:mac;
inj-event(amf_verify(x,y,z))==>
inj-event(ue_init(x,y,z)).
```

Confidentiality is a reachability property. Verifying reachability properties is the most basic capability of ProVerif. The tool can prove which terms are available to an attacker by checking all possible protocol executions and attacker behaviors. Using the following query statements, ProVerif can test the confidentiality of term *x*, such as the K_{AMF} and the uplink or downlink NAS overflow counter in the model.

```
query attacker(x).
```

3.3.2 Challenges and solutions in modeling the fast authentication procedure

We now model the fast authentication procedure in ProVerif, which takes 350 lines of ProVerif code. Due to a large amount of modeling code, we only describe the challenges encountered in the modeling process and the corresponding solutions.

Challenge 1: complex cryptographic primitives The fast authentication procedure involves complex encryption, integrity protection, and key derivation algorithms. Modeling these cryptographic primitives requires multiple mathematical operations, which is challenging.

```
fun senc_msg(bitstring,key):bitstring.
reduc forall x:bitstring,y:key;
sdec_msg(senc_msg(x,y),y)=x.
fun macBuilder(ies,container,kay,nat,
    nat):mac.
fun KDF_KNASenc(key):key.
fun KDF_KNASint(key):key.
```

Solution In our security assumptions, we consider these cryptographic primitives to be secure and public. Thus, we can ignore the complex details of cryptographic primitives. This does not affect analysis results. Specifically, we define the auxiliary cryptographic algorithms, including symmetric encryption, integrity protection, and key derivation. The encryption algorithm negotiated between the UE and the AMF is the symmetric encryption in essence. We choose the symmetric encryption and the integrity protection for the NAS signaling message. The key derivation is used to obtain encryption and integrity protection keys from K_{AMF}. Moreover, we simplify the input parameters of these primitives by considering only key parameters. For example, the 3GPP group stipulates that the key for encrypting the NAS signaling message needs to be derived from the K_{AMF} and the encryption algorithm type as parameters. However, the encryption algorithm type is known or enumerable to the attacker, so we simplify the cryptographic primitive to set the input parameter to only K_{AMF}.

Challenge 2: protection mechanism for the security context In the two storage scenarios, the security mechanisms proposed by 3GPP are different. When the security context is stored in the USIM card, the PIN-based access control should be modeled. And it is necessary to model the validity checking of the security context stored in the baseband chip. Moreover, the model should represent the fact that the security context stored in the baseband chip cannot be read by devices other than the baseband chip. However, it is very challenging to directly describe the above processing logic with ProVerif.

```
const UE_USIM:uetype.
const UE_BC:uetype.
fun pinContextMap(pin,context):map.
reduc forall x:pin,y:context;
    getContextByPin(pinContextMap(x,y),
    x)=y.
free s:bitstring[private].
fun supiContextMap(supi,context,
    bitstring):map.
reduc forall x:supi,y:context,z:
    bitstring;getContextBySupi(
    supiContextMap(x,y,z),x,z)=y.
```

Solution We define two types of UEs to distinguish different storage methods of the security context. We pass a UEType parameter to the UE process. If the value of this parameter is UE_USIM, the security context is stored in the USIM card. Otherwise, the security context is stored in the baseband chip. To indicate that the correct PIN code is required to read the security context stored in the USIM card, we define a map between the PIN and the security context. The UE can obtain the corresponding security context from the map through the PIN. We take the similar approach to model the scenario where the security context is stored in the baseband chip. Specifically, we define a map to represent the fact that the baseband chip stores the security context together with the SUPI. To indicate that only the baseband chip can read the context, we introduce a private parameter in the map, which only the baseband chip knows. The baseband chip can obtain the security context from the map through the SUPI and the private parameter.

Challenge 3: replay protection The AMF and the UE shall support replay protection based on the NAS COUNT for protected NAS messages. However, the 3GPP group does not define specific mechanism for implementing replay protection. Therefore, a secure implementation against replay attacks needs to be determined.

Solution Combining implementations of open-source projects and expert experience, we identify a secure implementation of replay protection. We define a type that represents the NAS COUNT, which can be split into two parts, namely the NAS overflow counter and the NAS SQN. We introduce the implementation of replay protection from the perspective of the AMF. And The UE process has a similar mechanism. Specifically, the AMF compares whether the received uplink NAS SQN is smaller than the local one. If this fact is true, the AMF will use the received uplink NAS SQN and a temporary value of the local uplink NAS overflow counter incremented by one as parameters to check the integrity protection. If the message integrity verifies correctly, the AMF update the local uplink NAS counter.

```
type count.
fun nasCt(nat,nat):count[data].
let amf= ...
  if UEUpSQN<AMFUpSQN then(
  let xMAC:mac=macBuilder(...,AMFUpOf
      +1,UEUpSQN)
  if xMAC=MAC then(
    let AMFUpOf=AMFUpOf+1 in
    let AMFUpSQN=UEUpSQN in ...
```

Table 1 Test results for security properties

Property	Type	UU	UB
C.	K_{AMF}	×	✓
	NAS COUNT	×	✓
A.	U.t.A	×	✓
	A.t.U	×	×

4 Security analysis

In this section, we first present the formal verification results of the fast authentication procedure based on the security context in 5G networks. We then expose two vulnerabilities that lead to the failure of security goals. Finally, we point out that these vulnerabilities still apply to 4G networks.

4.1 Formal analysis results in 5G networks

We depict the results of our analysis of the confidentiality properties and authentication properties as shown in Table 1. ‘C.’ means the confidentiality properties. ‘A.’ means the authentication properties. ‘UU.’ represents the security context stored in the USIM card. ‘UB.’ represents the security context stored in the baseband chip. ‘U. t. A.’ means that the UE can authenticate the AMF. ‘A. t. U.’ means that the AMF can authenticate the UE. The symbol ‘×’ denotes that the property is violated. The symbol ‘✓’ denotes that the property is met.

When the security context is stored in the USIM card, the confidentiality of K_{AMF} and the NAS COUNT cannot be guaranteed, which means that an attacker can obtain these two values. Moreover, the authentication properties between the UE and the AMF cannot be satisfied. In this scene, attackers can imitate the UE or the AMF to bypass the authentication of another entity. When the security context is stored in the baseband chip, the confidentiality of the K_{AMF} and the NAS COUNT cannot be compromised. In this scene, the UE can authenticate the AMF. However, the AMF cannot effectively authenticate the UE. An attacker can build a malicious UE entity to deceive the AMF.

4.2 Vulnerabilities

We now introduce two vulnerabilities that cause security goals to fail to meet expectations. An attacker could exploit these vulnerabilities to illegally use an effective security context of a victim. We analyze these vulnerabilities from the perspective of how the security context is stored.

The security context stored in the USIM card The 3GPP standard stipulates that the PIN-based access control can prevent the security context stored in the USIM card from being maliciously read (3GPP 2022b). However, the PIN-based access control cannot achieve the desired effect (Zhao et al. 2021). First, the USIM card cannot differentiate between various entities accessing the card through the PIN verification. To make matters worse, mobile carriers usually set a default PIN and disenable the PIN verification. Our survey results show that most subscribers also do not change the default PIN. Therefore, an attacker can access the security context saved in the USIM card, which uses the default PIN, with a card reader. In addition, the attacker can lure users to install a malicious application with the root privilege. Then the application can send APDUs to the USIM card to obtain the security context with the PIN access condition. In our work, we have tested USIM cards from three mobile carriers and found that they all use 1234 as the default PIN. Only one of the three mobile carriers has enhanced the USIM card to prevent the use of the card reader to obtain the security context. Mobile users of the remaining two mobile carriers may face various attacks, including eavesdropping, man-in-the-middle (MitM), and impersonation. Second, the channel between the baseband chip and the USIM card is vulnerable (3GPP 2022b). The attacker can use specific hardware to eavesdrop on the APDU traffic on the channel. When the baseband chip sends a security context to the USIM card, the attacker can obtain the unencrypted context without the PIN certification.

The security context stored in the baseband chip This protocol vulnerability is first pointed out in our work. The 3GPP standard stipulates that the security context stored in the baseband chip shall correspond to the subscriber's permanent identity. It also specifies that the security context stored in the baseband chip should be deleted from the baseband chip in the following cases (3GPP 2022e): (1) the USIM card is removed from the ME when the ME is in power on state; (2) the ME is turned on and finds that there is no USIM card in the ME; (3) the ME is turned on and discovers that the USIM card is different from the one which was used to create the security context. However, we find that if the ME is in airplane mode or turned off, the baseband chip cannot capture the USIM card being removed or replaced. The security context stored in the baseband chip can only be used if the permanent identity from the USIM matches the identity associated with the security context. Therefore, an attacker may construct a USIM card with the same permanent identity, and replace the legitimate USIM card with the fake USIM card when the phone is in airplane mode or turned off to illegally use the stored security context. However, the security context stored in the non-volatile memory of the baseband chip cannot be read by other software or hardware except the baseband chip, so the attacker cannot obtain some necessary information to launch the eavesdropping attack.

4.3 Evaluation in 4G networks

We evaluate whether the above vulnerabilities exist in the 4G network. Fortunately, the conclusion is that these vulnerabilities apply to 4G networks. We conduct analysis from the following perspectives, namely the security context, the key derivation, and the fast authentication procedure.

First, the 4G security context and the 5G security context have similar structures, storage methods, and security mechanisms. Specifically, the 4G security context includes the GUTI (ID: 6FE3) and the 4G NAS security context (ID: 6FE4). And the 4G NAS security context consists of the K_{ASME} with the associated key set identifier eKSI, the UE security capabilities, and a distinct pair of NAS COUNTs (one NAS COUNT for uplink and one NAS COUNT for downlink) (3GPP 2022a). Their functions are the same as the parameters in the 5G NAS security context. The UE and the Mobility Management Entity (MME) shall be capable of storing the 4G security context. In the UE, the 4G security context shall be stored in the USIM card (if supported by the USIM card) or in the non-volatile memory of the baseband chip (3GPP 2022d). The 4G security context stored in the USIM is also protected by a PIN-based access mechanism. And the 4G security context stored in the baseband chip can only be used if the IMSI from the USIM card matches the IMSI corresponding to the cached context.

Second, the key derivation in 4G networks is shown in Fig. 2. The keys for encryption and integrity protection are available from the K_{ASME}. In the process, the NAS COUNT should be used as the input parameter. The security context contains the important parameters needed for the key derivation.

Third, the 3GPP group defines the fast authentication procedure in 4G networks, which is similar in principle to the fast authentication procedure in 5G networks (3GPP 2022b, d). To put it simply, the UE sends a protected Attach Request message based on the security context to the MME. If the MME can find the corresponding security context according to the GUTI and the eKSI, and complete the verification of the message, the MME may not trigger a primary authentication and key agreement procedure.

Based on the above analysis, it can be concluded that the vulnerabilities uncovered in our work are applicable to the fast authentication procedure based on the security context in 4G networks. Our experiments in Sect. 5 also prove that these vulnerabilities do exist in 4G networks. In addition, the 4G security context and the 5G security context can be transformed into each other (3GPP 2022e). For example, during the change from the 4G network to the 5G network, if the UE has no valid 5G security context, the UE shall send the Registration Request message without integrity protection and encryption. And the message shall contain the Tracking Area Update (TAU) request integrity protected using the 4G security context shared with the MME. And the UE increments the uplink 4G NAS COUNT by one. Once receiving the message, the AMF shall interact with the MME identified by the GUTI included in the Registration Request message to retrieve the 4G security context. Specifically, the AMF sends the enclosed TAU request message to the MME. The MME then verifies the message using the stored 4G security context. If the verification is successful, the MME shall return the 4G security context to the AMF. Finally, the AMF and the UE create a fresh mapped 5G security context based on the 4G security context. This means that attackers can use the illegally obtained 4G security context to perform malicious actions in 5G networks. However, since the three mobile carriers involved in our work do not support this feature, we cannot conduct actual tests.

5 Basic attacks

In this section, we describe the proposed attacks that exploit the discovered vulnerabilities. First, we introduce the attack model and the experimental setup. We then detail two attacks, namely impersonation and eavesdropping. Table 2 shows the test results for three mobile carriers. We denote the three mobile carriers as C-I, C-II, and C-III for privacy concerns. ‘V1’ means that the PIN-based access control cannot prevent the security context from being illegally read. ‘V2’ means that the validity checking of the security context stored in the baseband chip can be bypassed. We found that the USIM cards assigned by these three carriers all support the storage of the 4G security context, but none of them support storing the 5G security context. Therefore, when accessing a 5G network, the security context will be stored in the non-volatile memory of the baseband chip. When accessing the 4G network, the security context is stored in the USIM card. Based on these facts, we perform different attack verifications in 4G and 5G networks. In addition, our analysis shows that when the security context is stored in the USIM card, these attacks can be effective for a long time after the one-time physical access. Finally, we introduce an informal survey result to evaluate the feasibility of these attack models.

5.1 Attack model

We assume an attacker can perform attacks in the following two scenarios.

In the first scenario, we assume the attacker can obtain the valid security context stored in the victim’s USIM card. This can be achieved with one-time physical access through hardware (e.g., the card reader) or malware installed on the victim’s phone. There are many types of malware, such as worms, backdoors, viruses, and Trojans, that easily threaten smartphones (Qamar et al. 2019) and can be employed by attackers to steal private information in potential. In our work, we take advantage of the card reader to obtain the target security context. In this case, we assume that the attacker knows the victim’s PIN or that the victim disables the PIN protection. Our survey shows that this is very likely to happen in real life. In addition, attackers can leverage the USIM sticker to extract sensitive files from the victim’s USIM card (Zhao et al. 2021). In this case, the attacker can bypass the PIN protection to obtain the security context. Our analysis shows that attackers can remotely launch long-term effective impersonation and eavesdropping attacks after the one-time physical access.

In the second scenario, we assume the attacker can obtain a mobile phone storing a valid security context of the victim. This is indeed achievable in real life. For example, on some special occasions, several people will take turns using a mobile phone, and each person will use a different USIM card. An attacker who is one of the members can then attack the previous user. Alternatively, the attacker can trick the victim into inserting their USIM card into the attacker’s phone in front of the victim and then return the USIM card. The attacker only needs to ensure that the phone is turned off or in airplane mode when the victim’s USIM card is pulled out. In this scenario, the attacker can launch the impersonation attack, but cannot eavesdrop on the victim’s traffic. Although the premise of the attack is relatively harsh, it is not completely impossible. Once the impersonation attack is successfully launched, the attacker can log into the victim’s application to steal privacy, which is described in Sect. 6.

Table 2 Test results for mobile carriers

Carrier	V1	V2	Attack			
			Impersonation	Eavesdropping	Location spoofing	OTAuth bypass
C-I	✓	✓	✓	✓	✓	✓
C-II	×	✓	✓	×	✓	✓
C-III	✓	✓	✓	✓	✓	✓

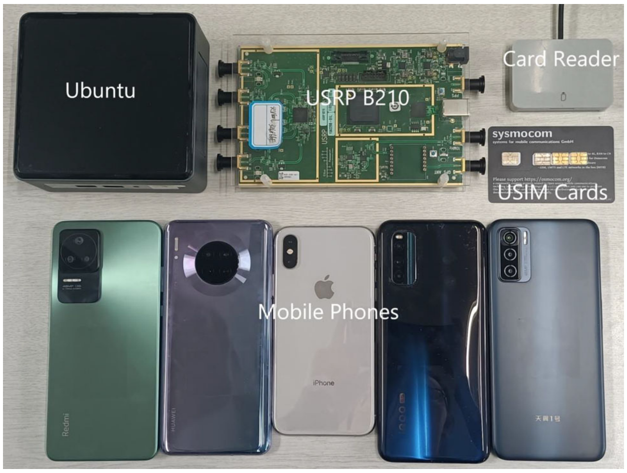


Fig. 3 The experimental hardware setup

Moreover, we assume that the attacker knows some basic knowledge about the 4G and 5G networks, and have the ability to use the card reader to read and write the USIM card. Attackers should have basic programming capabilities and can leverage Software-Defined Radio (SDR) devices. The attacker can set up a malicious cellular network to send messages to the UE.

5.2 Experimental setup

We present the main experimental hardware setup as shown in Fig. 3. One computer with the Ubuntu system runs the open-source srsLTE (Ismael et al. 2016) and the software that interacts with the USIM card through the card reader. The srsLTE is a 4G software radio suite, including a full-stack SDR 4G UE, a full-stack SDR 4G BS, and a lightweight 4G core network. And it needs to cooperate with the Universal Software Radio Peripheral (USRP) B210 (research 2022) to transmit and receive radio signals. In addition, our experiments involve 5 mobile phones from different baseband manufacturers as shown in Table 3. Among these mobile phones, only the iPhone XS equipped with the Intel baseband chip does not support the 5G network. And these mobile phones have vulnerabilities in judging the USIM status, which is described in Sect. 5.3. More seriously, we find that the TYH212U faces replay attacks due to the lack of sufficient checks on NAS COUNT. Details about this security threat are provided in Sect. 5.4.3. There are two kinds of USIM cards used in our experiments. One is the legitimate USIM card that is assigned by the mobile carrier and facilitates the subscriber to access the 4G and 5G network (we name it a real card). And the other is the programmable USIM card (we named it a fake card) (Sysmocom 2022). The files (e.g., IMSI, EPSLOCI, and EPSNSC) stored in the fake card can be read and written at will. None of these cards support storing the 5G security context.

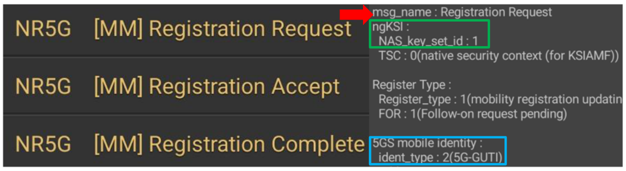
5.3 Impersonation using the security context in baseband chip

The security protection proposed by the 3GPP group for the security context stored in the baseband chip cannot achieve its intended purpose. First, when the mobile phone is in airplane mode or turned off, the baseband chip cannot determine whether the mobile phone card has been pulled out or replaced. None of the five mobile phones in our experiment can effectively judge the status of USIM as shown in Table 3. Second, the baseband decides whether to use the security context for the fast authentication procedure by judging whether the permanent identity in the USIM card is consistent with the identity corresponding to the saved security context. In our work, we propose an impersonation attack that exploits the security context stored in the baseband chip. We conduct this experiment in the mobile carriers' 5G networks.

Attack procedure First, the attacker determines the permanent identity SUPI of the victim's USIM card. Since these three carriers have not enabled protection against SUPI, SUPI is transmitted over the air interface in the form of IMSI at this time (Nie et al. 2022). Then the attacker writes it to the fake card. Second, the attacker would need to obtain a phone with a legitimate security context. It is indeed possible as described above. Third, the attacker inserts the fake card into the phone, and then turns off airplane mode or turns the phone on. Since the permanent identity in the fake card is consistent with the identity corresponding to the security context, the mobile phone determines that the saved security context is valid, and uses it to initiate a fast authentication procedure. Finally, the attacker can use the fake card to access the carrier's 5G network as shown in Fig. 4. In this registration procedure, the Registration Request message sent by the UE includes the ngKSI with the value of 1 (marked by the green

Table 3 Test results for mobile phones

Phone	Baseband	Support 5G?	USIM status?	Replay protection?
iPhone XS	Intel	×	×	✓
iQOO Z1	MediaTek	✓	×	✓
Huawei Mate30	Kirin	✓	×	✓
Redmi K40S	Qualcomm	✓	×	✓
TYH212U	Unisoc	✓	×	×



```
NR5G   [MM] Registration Request
NR5G   [MM] Registration Accept
NR5G   [MM] Registration Complete
```

```
msg_name : Registration Request
ngKSI :
  NAS_key_set_id : 1
  TSC : 0(native security context (for KSIAMF))

Register Type :
  Register_type : 1(mobility registration updatin
  FOR : 1(Follow-on request pending)

5GS mobile identity :
  ident_type : 2(5G-GUTI)
```

Fig. 4 The NAS signaling messages of the impersonation attack in a 5G network

box) and the GUTI indicating the mobile identity (marked by the blue box). And the AMF has skipped the AKA procedure.

Discussion The impersonation attack exploiting the novel protocol vulnerability is applicable to the three mobile carriers that support the fast authentication procedure. The attacker needs to ensure that the victim is no longer connected to the network during the attack. Otherwise, the security context saved in the phone will be invalid. However, attackers can launch the OTAuth bypass attack discussed below within seconds of accessing the network. Thereafter, although the victim will reconnect to the network, which ends the impersonation attack, the attacker was already logged into the victim's application account at this point. In addition, as long as mobile users do not switch to the USIM card that supports storing the 5G security context, they will continue to face the security threat caused by the impersonation attack. More seriously, even if the victim enables the PIN protection and does not use the default PIN, the attack is still possible.

5.4 Attacks using the security context in USIM card

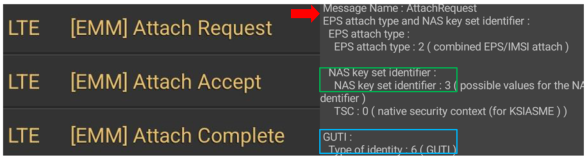
The PIN-based access control cannot effectively protect the security context stored in the USIM card. If the mobile carrier does not enhance the security of the USIM card, an attacker can use hardware or software to obtain these security contexts through one-time physical access to the victim's mobile phone. We find that the vulnerability poses real threats to subscribers of C-I and C-III as shown in Table 2. C-II has made security enhancements to the USIM card to prevent malicious reading of the security context. To force the mobile phone to store the security context into the USIM card, the attacker can trigger the mobile phone to initiate the deregistration procedure by turning off the phone or enabling airplane mode.

The attacker pulls out the USIM card and reads the security context with the card reader. The attacker can then insert the USIM card back into the victim's phone. The process goes unnoticed by the victim. The attacker with a valid security context can perform several attacks (Zhao et al. 2021), such as traffic eavesdropping, impersonation, and MitM. The MitM attack requires the attacker to set up a malicious base station and core network with the same network parameters as the victim's mobile carrier. This fake network uses the security context to derive various keys to deceive the victim. The victim connects to the Internet through the fake network. Then the attacker can perform many MitM attacks such as IP/DNS spoofing and SSL hijacking. In this paper, we detail the improved impersonation and traffic eavesdropping attacks. Moreover, we further analyze that the attacker can continue to launch these attacks for a long period of time, even if the attacker is far away from the attacker. Since the mobile carriers' USIM cards cannot store the 5G security context, we validate these attacks in 4G networks. These attacks will be possible in 5G networks when carriers upgrade USIM cards to support storing 5G security context.

5.4.1 Impersonation

The attacker with a valid security context can impersonate a victim to access the network. The attack can be launched remotely from the victim.

Attack procedure We assume that the victim has connected to the mobile carrier's 4G network using a mobile phone equipped with a real card. And a valid 4G security context is already stored in the real card. The attacker can then launch the impersonation attack. First, the attacker uses the card reader to obtain the files (i.e., IMSI, EPSLOCI, and EPSNSC) required for the fast authentication procedure. Second, the attacker uses the same method to write these files into the fake card. Then the attacker inserts the fake card into the phone and uses the fast authentication procedure to access the mobile carrier's 4G network as the victim. Figure 5 shows the NAS signaling messages. The Attach Request message includes the eKSI (marked by the green box) and the GUTI (marked by the blue box), and the 4G network does not initiate an AKA procedure. Both sides use the key in file EPSNSC to derive various keys to ensure the security of subsequent communications.



```
LTE    [EMM] Attach Request      Message Name : AttachRequest
                                 EPS attach type and NAS key set identifier :
                                   EPS attach type :
                                     EPS attach type : 2 ( combined EPS/IMSI attach )
LTE    [EMM] Attach Accept
                                   NAS key set identifier :
                                     NAS key set identifier : 3 ( possible values for the NA
                                 dentifier )
                                     TSC : 0 ( native security context (for KSIASME ) )
LTE    [EMM] Attach Complete
                                 GUTI :
                                   Type of identity : 6 ( GUTI )
```

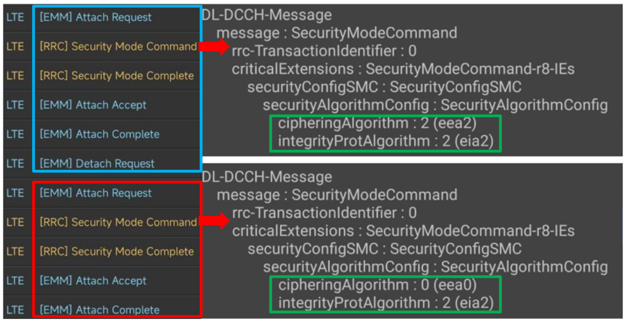
Fig. 5 The NAS signaling messages of the impersonation attack in a 4G network

Discussion We found that the network sometimes initiates the identification procedure. Therefore, we set the IMSI of the fake card as the permanent identity of the victim to be able to properly handle the identity requests initiated by the network. The impersonation attack proposed by Zhao et al. (2021) is implemented using the srsLTE. However, since the software utilizes the SDR to send wireless signals, it is not as stable as mobile phones equipped with a baseband chip. In our work, the impersonation attack based on this vulnerability is performed with a commercial mobile phone. Moreover, there are various applications on the mobile phone, and the attacker may be able to use them to launch further attacks, such as the OTAuth bypass.

5.4.2 Eavesdropping

The attacker with a valid security context can impersonate a legitimate network to eavesdrop on the user traffic. The attacker needs to be close to the victim to guarantee the victim's access to the established malicious network.

Attack procedure We assume that the attacker has obtained a valid security context of the victim, and the victim chooses to access the 4G network. First, the attacker modifies the core network code of srsLTE and hardcodes the security context into a temporary variable. The attacker should configure the parameters of srsLTE to be consistent with the 4G network of the target mobile carrier. These parameters include the Public Land Mobile Network (PLMN), the Tracking Area Code (TAC), the Access Point Name (APN), and the downlink frequency. The attacker increases the transmit power of the USRP B210 as the front end so that the malicious network with the strongest signal is accessed by the victim. The malicious network then leverages the security context to derive various keys needed for subsequent communications. The signaling messages of the victim's two attach procedures based on the same security context are shown in Fig. 6. In the first procedure (marked by the blue box), the victim accesses the mobile carrier's 4G network, which uses the encryption algorithm to protect the data on the air interface. In the second procedure (marked by the red box), the victim connects to the malicious network, which does not encrypt the data on the air interface.



```
LTE  [EMM] Attach Request            DL-DCCH-Message
LTE  [RRC] Security Mode Command       message : SecurityModeCommand
LTE  [RRC] Security Mode Complete        rrc-TransactionIdentifier : 0
LTE  [EMM] Attach Accept                 criticalExtensions : SecurityModeCommand-r8-IEs
LTE  [EMM] Attach Complete                 securityConfigSMC : SecurityConfigSMC
LTE  [EMM] Detach Request                    securityAlgorithmConfig : SecurityAlgorithmConfig
                                               cipheringAlgorithm : 2 (eea2)
                                               integrityProtAlgorithm : 2 (eia2)
LTE  [EMM] Attach Request            DL-DCCH-Message
LTE  [RRC] Security Mode Command       message : SecurityModeCommand
LTE  [RRC] Security Mode Complete        rrc-TransactionIdentifier : 0
LTE  [EMM] Attach Accept                 criticalExtensions : SecurityModeCommand-r8-IEs
LTE  [EMM] Attach Complete                 securityConfigSMC : SecurityConfigSMC
                                             securityAlgorithmConfig : SecurityAlgorithmConfig
                                               cipheringAlgorithm : 0 (eea0)
                                               integrityProtAlgorithm : 2 (eia2)
```

Fig. 6 The signaling messages of two attach procedures based on the same security context

Discussion The encryption key used to protect the data transmitted over the air interface is derived by the core network based on the NAS security context. The UE and the base station negotiate the encryption algorithm through the security mode control procedure of the Radio Resource Control (RRC) protocol. Changing the encryption algorithm chosen by the base station does not invalidate the security context. Therefore, we configure the malicious base station to use the null encryption algorithm. Such a trick can avoid the complicated decryption process when listening to user traffic (Zhao et al. 2021). In addition, the attacker can further launch the DNS spoofing attack, which redirects the victim to a malicious website (Rupprecht et al. 2019). The attacker can then steal the victim's username and password.

5.4.3 Persistence of attacks using the security context in USIM card

To launch the attacks (including the impersonation and the eavesdropping) using the security context stored in the USIM card, the attacker needs to perform the one-time physical access with the USIM card. However, we found that these attacks based on the valid security context acquired by this physical access can be effective for a long time.

Except for the NAS COUNT value that is incremented with the transmission of NAS signaling messages, other data of the security context remains unchanged until a new AKA procedure occurs. The new AKA procedure will happen in the following two situations: (1) the network actively initiates the AKA procedure. We connect the five phones to the three mobile carriers for 24 h and find that none of the three carriers will actively initiate the AKA process. (2) The NAS COUNT value wraps around (close to 2^{24}). However, this rarely happens due to the low exchange of NAS signaling messages. We counted the number of NAS signaling messages transmitted by the UE in one hour after accessing the 4G and 5G networks of C-I and C-III in the two scenarios of enabling the cellular data and disabling the cellular data. During the testing, the mobile phone follows two daily usage scenarios: static state and mobile state. The results are shown in Table 4. When the cellular data is turned on, the UE will initiate a ser-

Table 4 Statistics on the number of NAS signaling messages transmitted by the UE in two scenarios

Carrier	5G				4G			
	Static		Mobile		Static		Mobile	
	Enable	Disable	Enable	Disable	Enable	Disable	Enable	Disable
C-I	138	6	37	8	131	15	16	6
C-III	114	7	29	9	141	8	20	8

vice request from time to time, which leads to an increase in the number of NAS signaling messages. Moreover, when the mobile phone is in the mobile state, the continuous switching of the RRC connection makes the mobile phone maintain an active connection with the network for a longer period of time. Therefore, the number of service request messages initiated by the phone in the mobile state is less than that initiated by the phone in the static state. Either way, it would take 13 years for the NAS COUNT value to wrap around under the ideal condition.

Therefore, the attacker only needs to deduce the latest NAS COUNT values to launch these attacks again after a period of time. We have analyzed the processing specification of the NAS COUNT value stipulated by the 3GPP group (3GPP 2022c, d). The sender shall use its locally stored 24-bit NAS COUNT values as input to the ciphering and the integrity protection algorithm. The receiver shall use the 8-bit NAS SQN included in the received message and an estimate for the 16-bit NAS overflow counter. An attacker can infer the NAS COUNT value by establishing a malicious network and a malicious UE. Specifically, after the victim accesses the malicious network, the attacker can obtain the NAS SQN included in the message sent by the UE. Then, the attacker can continuously check the integrity of the message by incrementing the local NAS overflow counter until it succeeds. Experiments show that it only takes 0.5s to perform an integrity check. If the UE does not receive a response after sending the message for 15 s, it will send a second request. The UE will try this procedure up to 5 times. Therefore, the attacker has enough time to deduce the NAS overflow counter. For example, the attacker launches these attacks again one day after obtaining the security context. According to our statistics, the UE may transmit about 3000 NAS signaling messages during this period, and the attacker only needs to perform 12 (3000/256) attempts to get the NAS overflow counter. After getting the latest uplink NAS COUNT value, the attacker can use the malicious UE to send legitimate uplink messages. After receiving the legitimate downlink message returned by the mobile carrier, the attacker can use the same method to infer the downlink NAS overflow counter. It is worth noting that before the end of the attack, the attacker needs to send a specific number of NAS signaling messages to keep the NAS COUNT values jointly maintained between the victim and the mobile carrier in sync.

The above analysis shows that these attacks exploiting the security context stored in the USIM card are valid for extended periods of time. Furthermore, we found that the three carriers update the GUTI every time the UE connects to the network. Therefore, the attacker needs to obtain the updated GUTI of the victim. Whenever a UE sends a message containing the GUTI to access the malicious network established by an attacker, the attacker can send an unprotected identity request message to the UE. Then the UE will return an identity response message containing the IMSI. This message is integrity protected by the security context, and this procedure will not cause the UE to delete the security context. The attacker can determine the latest GUTI of the victim based on the target IMSI. In addition, the attacker can also obtain the victim's GUTI by sniffing the paging messages (Hong et al. 2018).

During the experiment, we also found some insecure implementations about handling the NAS COUNT value: (1) network. The core network of the open-source project srsLTE (Ismael et al. 2016) does not check whether the NAS COUNT value is replayed. The open-source core network free5GC (Free5GC 2022) and open5GS (Open5GS 2022) update the local NAS COUNT value before the integrity verification is successful. An attacker can exploit this flaw to desynchronize the NAS COUNT values between the UE and the network, thereby causing the DoS attack on the UE. (2) UE. The commercial mobile phone TYH212U equipped with the baseband chip from the Unisoc lacks an effective replay protection check for the NAS COUNT value as shown in Table 3. This leads to the replay attack of NAS signaling messages.

5.5 A survey

To evaluate the feasibility of these attacks discussed in our work, we have conducted an informal survey and obtained responses from 147 volunteers. As shown in Fig. 7, the results show that these attacks are feasible in practice. Specifically, 95.92% of the participants set the password for their devices, but only 25.17% of the users enable the PIN verification for their USIM cards. To make matters worse, among the users who enable the PIN verification, 83.67% of the subscribers use the default PIN. The security contexts stored in the USIM cards of these users may be exposed to attackers. In addition, 59.18% of subscribers are still using 4G USIM cards, and

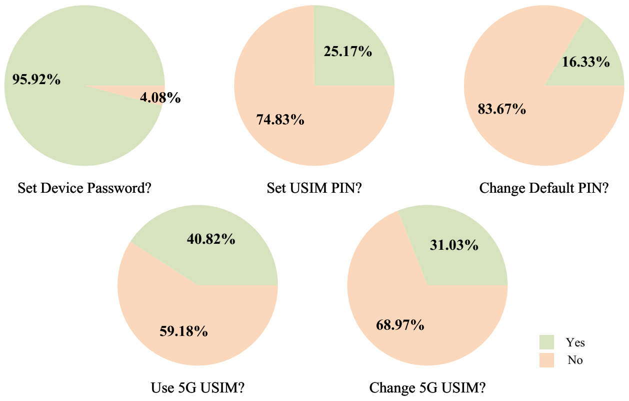


Fig. 7 The USIM usage survey

68.97% of them are unwilling to upgrade to 5G USIM cards. Therefore, when they access the 5G network, the 5G security context is stored in the mobile phone. This may be exploited by attackers to launch the impersonation attack.

6 Further security threats caused by impersonation

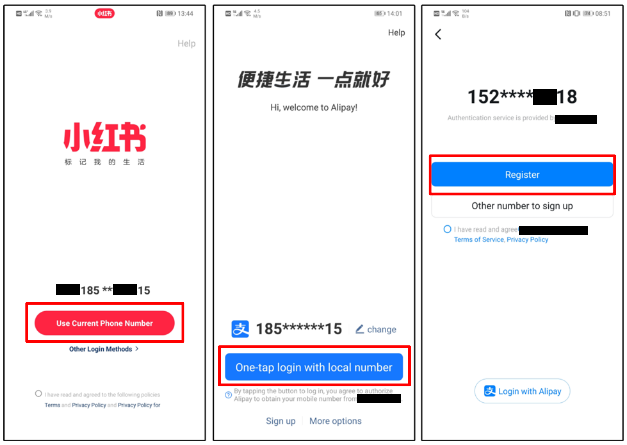
In this section, we introduce two security threats caused by the impersonation attack, namely location spoofing and one-tap authentication bypass.

6.1 Location spoofing

The mobile carrier can locate a subscriber based on which base station the subscriber is connected to. However, the attacker can launch impersonation attacks far from the victim. Since the geographical location of the base station accessed by the attacker is different from the real geographical location of the victim, this attack can pose a challenge for the mobile carrier to locate the subscriber. Although the network disconnects the attacker when the victim reconnects to the network, the location spoofing attack works well in the scenario where the attacker is the victim. For example, if the vulnerability is exploited by criminals to forge their locations, they could fabricate evidence and evade the law.

6.2 One-tap authentication bypass

Almost all apps support the SMS-based authentication. To receive the verification code contained in one SMS, the attacker needs to access the IP Multimedia Subsystem (IMS) network, which forces an AKA procedure. Therefore, the attacker cannot bypass the SMS-based authentication through the impersonation attack. However, the emerging OTAuth service gives attackers the possibility to log into the victim's account. The OTAuth service provided by the mobile carrier allows users to quickly sign up or log in to their application accounts with the local phone number of the mobile phone with just one tap on the button as shown in Fig. 8 (marked by the red box) (Zhou et al. 2022). All three mobile carriers discussed in our work support this service. Nearly 400 popular apps have embedded OTAuth services, such as Alipay, Weibo, TikTok, etc.



○ I have read and agreed to the following policies
Terms and Privacy Policy and Privacy Policy for

By tapping the button to log in, you agree to authorize Alipay to obtain your mobile number from [redacted].

Authentication service is provided by [redacted]

○ I have read and agree [redacted]
Terms of Service, Privacy Policy

Fig. 8 Examples of OTAuth interfaces in three apps provided by mobile carriers

The core part of the service is that the local phone number is not obtained through any system functions, but based on the mobile carrier's unique capability to identify the phone number. The principle of this service is shown in Fig. 9. Specifically, the mobile phone equipped with a valid USIM card accesses the mobile carrier's 4G or 5G network through the AKA procedure or the fast authentication based on the security context. After going through the security mode control procedure, the carrier will assign a Local Area Network (LAN) IP address to the phone. When the app client that supports the OTAuth service initiates an OTAuth request through the 4G/5G network, the carrier can recognize the phone number based on the IP address and return a token bound to the phone number. The app client then forwards the token to the app server. The app server uses this token to request the user's phone number from the carrier. Finally, the app server decides whether to allow the login or registration request according to the phone number.

The requirement for the OTAuth service is to exchange data through the 4G/5G network. Therefore, the attacker can use the impersonation attack to bypass the authentication scheme to log in to the victim's account. In addition, we found that some apps that support the OTAuth service perform additional verification on new devices, such as the SMS-base authentication or requiring the full phone number. The test results for 20 popular apps supporting the OTAuth service are shown in Table 5. Among these apps, only 3 apps (i.e., Alipay, QQ, and Douyu TV) authenticate new devices. Further, inspired by the fact that some apps support logging

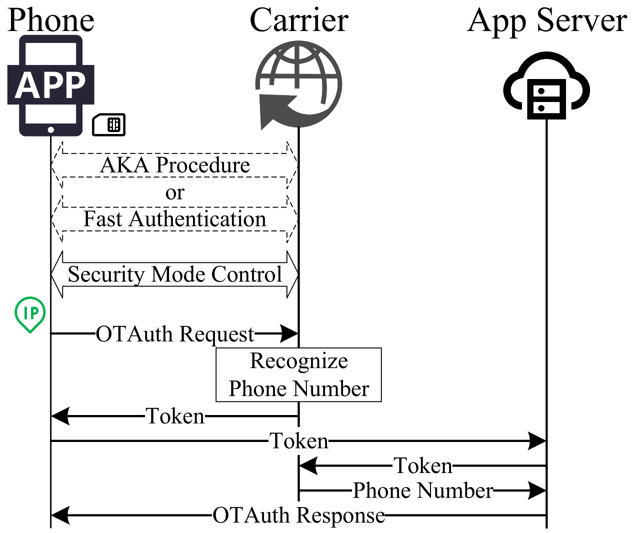


Fig. 9 The principle of the OTAuth service

Table 5 Test results on 20 popular apps

App	New device?	App	New device?
Alipay	×	Kuaishou	✓
QQ	×	Toutiao	✓
Taobao	✓	Weibo	✓
TikTok	✓	Bilibili	✓
Baidu	✓	Jingdong	✓
Ctrip	✓	Iqiyi	✓
Meituan	✓	Dianping	✓
Youku	✓	Gaode Map	✓
Meitu	✓	WPS Office	✓
Zhihu	✓	Douyu TV	×

in with other app accounts, our attack can affect more apps. An attacker can bypass authentication to log in to an app that supports the OTAuth service, and then use the account to log in to other apps that do not support the OTAuth service. For example, an attacker can use a Weibo account to log in to UC Browser, which does not support the OTAuth service.

7 Countermeasures

The vulnerabilities stem from the design flaw of the 3GPP protocol and the unsafe practices by mobile carriers. We propose the following countermeasures from four perspectives.

The mobile carrier First, the carrier should enhance the security of the USIM card to prevent the security context saved in the USIM card from being illegally read. Second, the carrier can disable the fast authentication procedure based on the security context. Moreover, for the UE that has already accessed the network, it is better for the mobile carrier to reinitiate an AKA procedure at regular intervals to update the security context. Third, the 5G USIM card should be enhanced to support storing the 5G security context and concealing SUPI.

The user On the one hand, users can enable PIN verification and modify the default PIN code at the same time. On the other hand, after the operator enhances the 5G USIM card, it is better for users to upgrade to the latest 5G USIM card. Not only does this prevent the security context from being stored in the baseband chip, but it also leverages the security features of 5G to protect its permanent identity.

The 3GPP protocol First, checking the validity of the security context stored in the baseband chip should incorporate other parameters (e.g., ICCID), not just the permanent identity. This will greatly increase the difficulty of the attack. Second, the Public Key Infrastructure (PKI) can be considered to join the existing USIM card security standard (Zhao et al. 2021). Third, the 3GPP group should specify the processing standard for the NAS COUNT to avoid possible security problems in the implementation of various manufacturers.

The equipment manufacturer Mobile phone manufacturers and baseband chip manufacturers should coordinate the transmission of USIM card status information. When the mobile phone is in airplane mode or turned off, the baseband chip should be able to know the change of the USIM card status information, and then decide whether to delete the security context.

8 Related work

We discuss the related work with the following two categories, namely, attacks in mobile networks, and security research on the USIM card.

Attacks in mobile networks It is well known that the authentication vulnerabilities in 2G networks allow attackers to perform impersonation, eavesdropping, and MitM attack (Zhang et al. 2020). Meyer and Wetzel (2004) exploit the signaling message lacking integrity protection in 3G networks to launch the MitM attack. The authentication relay attack in 4G networks allows an attacker to connect to the network without nullifying the security capabilities (Hussain et al. 2018). But the attacker cannot decrypt or send encrypted messages. Rupprecht et al. (2020) propose an impersonation attack that exploits the lack of integrity protection on the user plane in 4G networks and the reflection mechanism of Internet Control Message Protocol (ICMP). Zheng et al. (2017) find one vulnerability in Circuit Switched Fallback (CSFB) where the authentication step is missing, which allows an attacker to impersonate a victim. The downgrade attack can be exploited to obtain the victim's phone number and launch the MitM attack (Yu et al. 2019). Insecure implementations in the IMS system can be used to spoof SMS (Tu et al. 2016). Chen et al. (2023) utilize natural language processing and machine learning to build test cases from 3GPP specification and find several insecure handlings of the 4G security context on 22 popular mobile phones. These vulnerabilities expose these phones to serious security risks such as reply attacks.

Security research on the USIM card The USIM card cloning attack using the side-channel information was demonstrated (Liu et al. 2015). The SIMjacker has been exploited to perform many attacks, such as denial of service, scam calls, and information leakage (Security 2022). To find configuration bugs in USIM cards, the open-source tool SIMtester is developed (Srlabs 2022). Chitroub et al. (2018) study the security mechanisms of the embedded USIM (eSIM) and arise the potential vulnerabilities, which can be exploited to exhaust the memory of eSIM. Zhao et al. (2021) uncover three vulnerabilities of the PIN-based access control of the USIM card.

9 Conclusion

In this paper, we systematically formalize the fast authentication procedure based on the security context in 5G networks. The analysis shows that the existing security mechanisms cannot achieve the expected security goals. We identify two vulnerabilities, including one that has not been reported before. Further analysis revealed that these vulnerabilities also apply to 4G networks. Exploiting these vulnerabilities, we propose two types of attacks, namely impersonation and eavesdropping. In three mobile carriers, we have verified these attacks with five mobile phones from different baseband manufacturers. For one of these mobile phones and several open-source projects, we found some insecure implementations of handling the security context. In addition, we further analyzed the security threats caused by these attacks. The most serious is that an attacker can impersonate the victim to log in to multiple app accounts. The security issue involves more than 400 apps. In addition, we have conducted a detailed analysis of the persistence of the attacks, which can last for a long time. Finally, we propose several countermeasures from multiple perspectives to defend against these attacks. We have reported these vulnerabilities to the relevant authorized parties and received their confirmation.

Acknowledgements This work is supported by the National Natural Science Foundation of China (No. 62001055 and 61872386), and the Beijing University of Posts and Telecommunications-China Mobile Research Institute Joint Innovation Center.

Funding The authors have not disclosed any funding.

Data availability Enquiries about data availability should be directed to the authors.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

3GPP (2018) Study on authentication enhancements in 5g system (5gs). 3gpp tr 33.846. https://www.3gpp.org/dynareport/33846.htm

3GPP (2022a) 3gpp system architecture evolution (sae); security architecture. 3gpp ts 33.401. https://www.3gpp.org/dynareport/33401.htm

3GPP (2022b) Characteristics of the universal subscriber identity module (usim) application. https://www.3gpp.org/dynareport/31102.htm

3GPP (2022c) Non-access-stratum (nas) protocol for 5g system (5gs). 3gpp ts 24.501. https://www.3gpp.org/dynareport/24501.htm

3GPP (2022d) Non-access-stratum (nas) protocol for evolved packet system (eps). 3gpp ts 24.301. https://www.3gpp.org/dynareport/24301.htm

3GPP (2022e) Security architecture and procedures for 5g system. 3gpp ts 33.501. https://www.3gpp.org/dynareport/33501.htm

Basin D, Dreier J, Hirschi L et al (2018) A formal analysis of 5g authentication. In: Proceedings of the 2018 ACM SIGSAC conference on computer and communications security. Association for Computing Machinery, New York, NY, USA, pp 1383–1396

Beurdouche B, Bhargavan K, Delignat-Lavaud A et al (2015) A messy state of the union: taming the composite state machines of tls. In: 2015 IEEE symposium on security and privacy, pp 535–552

Blanchet B, Smyth B, Cheval V et al (2018) Proverif 2.04: automatic cryptographic protocol verifier, user manual and tutorial. https://bblanche.gitlabpages.inria.fr/proverif/manual.pdf

Borgaonkar R, Hirschi L, Park S et al (2019) New privacy threat on 3g, 4g, and upcoming 5g aka protocols. In: Proceedings on privacy enhancing technologies, pp 108–127

Chen Y, Tang D, Yao Y et al (2023) Sherlock on specs: building lte conformance tests through automated reasoning. In: USENIX security symposium

Chitroub S, Zidouni N, Aouadia H et al (2018) Sim card of the next-generation wireless networks: security, potential vulnerabilities and solutions. In: 2018 2nd European conference on electrical engineering and computer science, pp 502–509

Cui Z, Cui B, Su L et al (2023) Attacks against security context in 5g network. https://arxiv.org/abs/2303.10955

Free5GC (2022) what is free5gc. https://www.free5gc.org/

GSMA (2022) Global mobile trends 2023: navigating an uncertain world. https://data.gsmaintelligence.com/research/research/research-2023/global-mobile-trends-2023

GSMA (2023) Gsma mobile security research acknowledgments. https://www.gsma.com/security/gsma-mobile-security-research-acknowledgements/

Hong B, Bae S, Kim Y (2018) Guti reallocation demystified: Cellular location tracking with changing temporary identifier. In: Network and distributed system security symposium

Hussain S, Chowdhury O, Mehnaz S et al (2018) Lteinspector: a systematic approach for adversarial testing of 4g lte. In: Network and distributed system security symposium

Ismael GM, Andres GS, Sutton PD, et al (2016) Srslte: an open-source platform for lte evolution and experimentation. In: Proceedings of the tenth ACM international workshop on wireless network testbeds, experimental evaluation, and characterization. Association for Computing Machinery, New York, NY, USA, pp 25–32

Liu J, Yu Y, Standaert F et al (2015) Small tweaks do not help: differential power analysis of milenage implementations in 3g/4g usim cards. Computer security—ESORICS 2015. Springer, Cham, pp 468–480

Lowe G (1997) A hierarchy of authentication specifications. In: Proceedings 10th computer security foundations workshop, pp 31–43

Meyer U, Wetzel S (2004) A man-in-the-middle attack on umts. In: Proceedings of the 3rd ACM workshop on wireless security, pp 90–97

Nie S, Zhang Y, Wan T et al (2022) Measuring the deployment of 5g security enhancement. In: Proceedings of the 15th ACM conference on security and privacy in wireless and mobile networks. Association for Computing Machinery, New York, NY, USA, pp 169–174

Open5GS (2022) Open5gs. https://open5gs.org/

Qamar A, Karim A, Chang V (2019) Mobile malware attacks: review, taxonomy & future directions. Future Gener Comput Syst 97:887–909

research E (2022) Usrp b210. https://www.ettus.com/all-products/ub210-kit/

Rupprecht D, Kohls K, Holz T et al (2019) Breaking lte on layer two. In: 2019 IEEE symposium on security and privacy, pp 1121–1136

Rupprecht D, Kohls K, Holz T et al (2020) Imp4gt: impersonation attacks in 4g networks. In: Network and distributed system security symposium

Security A (2022) new simjacker vulnerability exploited by surveillance companies for espionage operation. https://simjacker.com/

Shaik A, Seifert J, Borgaonkar R et al (2016) Practical attacks against privacy and availability in 4g/lte mobile communication systems. In: Network and distributed system security symposium

Shaik A, Borgaonkar R, Park S et al (2019) New vulnerabilities in 4g and 5g cellular access network protocols: exposing device capabilities. In: Proceedings of the 12th conference on security and privacy in wireless and mobile networks. Association for Computing Machinery, New York, NY, USA, pp 221–231

Srlabs (2022) Simtester. https://github.com/srlabs/SIMtester

Sysmocom (2022) sysmocom usim cards. https://sysmocom.de/manuals/sysmousim-manual.pdf

Tu G, Li C, Peng C et al (2016) New security threats caused by ims-based sms service in 4g lte networks. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. Association for Computing Machinery, New York, NY, USA, pp 1118–1130

Unisoc (2023) Product security acknowledgements. https://www.unisoc.com/en_us/secy/thank

Wang Y, Zhang Z, Xie Y (2021) Privacy-preserving and standard-compatible aka protocol for 5g. In: USENIX security symposium, pp 3595–3612

Yu C, Chen S, Cai Z (2019) Lte phone number catcher: a practical attack against mobile privacy. Secur Commun Netw 2019:1–10

Zhang Y, Liu B, Lu C et al (2020) Lies in the air: characterizing fake-base-station spam ecosystem in china. In: Proceedings of the 2020 ACM SIGSAC conference on computer and communications security. Association for Computing Machinery, New York, NY, USA, pp 521–534

Zhao J, Ding B, Guo Y et al (2021) Securesim: rethinking authentication and access control for sim/esim. In: Proceedings of the 27th annual international conference on mobile computing and networking. Association for Computing Machinery, New York, NY, USA, pp 451–464

Zheng Y, Huang L, Shan H et al (2017) Ghost telephonist impersonates you: vulnerability in 4g lte cs fallback. In: 2017 IEEE conference on communications and network security, pp 1–9

Zhou Z, Han X, Chen Z, et al (2022) Simulation: demystifying (insecure) cellular network based one-tap authentication services. In: 2022 52nd annual IEEE/IFIP international conference on dependable systems and networks, pp 534–546

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