

It's not a bug, it's a feature: 25 years of mobile network insecurity *

Audun Jøsang¹, Laurent Miralabé² and Léonard Dallot²

¹ University of Oslo, Norway

² TazTag, France

josang@ifi.uio.no, lm@taztag.com, leonard.dallot@taztag.com

Abstract:

The global mobile networks are built with a set of core technologies developed during the 1980s, combined with subsequent generations of networking technologies for improved performance. Due to political pressure by European governments during the 1980 the initial GSM network, commonly called 2G, was purposely designed and built with weak security to allow easy interception of phone traffic by law enforcement agencies. Despite strengthened security in the more recent networking technologies of 3G and 4G, the weak security of 2G represents the 'weakest link' which thereby limits the security level of mobile networks in general. While this cybersecurity vulnerability is currently exploited by domestic law enforcement agencies for legal interception and surveillance, as well as by criminal and foreign powers for cybercrime and espionage, it is interesting to notice that it was created on purpose. This paper describes the background and the evolution of mobile network security, analyses the nature and consequences of security vulnerabilities in mobile networks, and proposes political and technical solutions to mitigate the threats posed by these vulnerabilities.

Keywords: Cyber security, surveillance, security politics, GSM, mobile network, SIM, IMSI, 2G, 3G, 4G.

1. Introduction

The digital cellular mobile network GSM (Global System Mobile), commonly called 2G, was standardised by ETSI (European Telecommunications Standards Institute) during the 1980s. Weak security was purposely built into the system because various European governments requested the ability to deactivate or break the encryption on the radio link in order to eavesdrop on mobile phone conversations. At the introduction of 2G GSM in 1991 MNOs (Mobile Network Operators) outside Europe were forced to use weak encryption, whereas European operators could use relatively strong encryption. For that purpose a strong and a weak set of cryptographic algorithms were designed. Parts of the GSM 2G standards were kept confidential and only distributed to industry partners under non-disclosure agreements. Mobile handsets for 2G were designed to handle both weak and strong encryption so they could be used anywhere in the world. An unfortunate consequence of this design is that by setting up fake base stations, an attacker can trick mobile phones to use weak or no encryption, even in countries where 2G operators use strong encryption. This is possible because network authentication was not included in the 2G standard, with the consequence that mobile phones do not know whether they are connected to a genuine 2G operator's base station, or to a fake base station set up by an attacker. This vulnerability enables attackers to break the encryption for any subscriber in a radio cell, obtain the encryption key, and then later eavesdrop on that subscriber, even, under certain conditions, if strong encryption is being used.

The subsequent UMTS (Universal Mobile Telecommunications System) networks, commonly called 3G, was standardised by 3GPP¹ under the responsibility of ITU-T² and launched around 2000. Criticism of the weak security in 2G resulted in relatively stronger mobile security being designed for 3G, e.g. with stronger cryptographic algorithms and a form of network authentication. However, 3G stops short of allowing the mobile phone/user to actually authenticate the identity of the mobile network.

Then around 2010, the latest mobile network technology named LTE (Long Term Evolution), commonly called 4G, was launched with additional security improvements.

In today's mobile networks a combination of 2G, 3G and 4G are being used worldwide. Mobile phones sold today are designed so that they can connect to all these networks in order for the phone to get maximum coverage most places. It is then obvious that 2G is the weakest link in the mobile network security chain. Although a mobile phone is able to communicate securely over 3G or 4G, the phone can simply be tricked not

* Proceedings of the 14th European Conference on Cyber Warfare and Security (ECCWS-2015). Hatfield, UK, July 2015

¹ 3rd Generation Partnership Project

² ITU-T: The International Telecommunication Union, Telecommunication Standardization Sector

to do so. In that sense the typical smartphone is rather dumb. It is as if we were tricked to secure the front door with only a hook, even if the door is equipped with an unpickable lock.

When searching for the reason behind the vulnerabilities in mobile networks it is interesting to notice that they are partially created on purpose by national and industry policy. It was politically desirable to create weak mobile security in 2G during the 1980s. Even if the political reasons no longer exist today there are now business incentives for keeping 2G and its weak security in operation. It must have been obvious to the designers of 3G and 4G that as long as 2G is still being used, making 3G and 4G more secure did not really improve the overall security. This unfortunate situation of mobile network insecurity is described in the subsequent sections, and in the discussion we propose solutions to mitigate the current vulnerabilities.

2. Mobile Network Security

2.1. Security in 2G GSM

The first time a MS (Mobile Station) enters the coverage area of a MNO (Mobile Network Operator) and requests registration, the base station first requests the permanent IMSI (International Mobile Subscriber Identity) in order to identify the subscriber. Subsequently, a short-lived TMSI (Temporary Mobile Subscriber Identity) is normally sent to the BS (Base Station). The IMSI is sensitive information because it can be used to track the subscriber. The purpose of sending the TMSI instead most of the time is precisely to minimise exposure of the IMSI. However, a BS has the possibility to request the IMSI at any time, which undermines the purpose of the TMSI, and which in fact is the vulnerability exploited by so-called “IMSI catchers”.

Authentication and encryption in 2G GSM are facilitated by a long-term secret 128 bit individual subscriber key K_i which is stored within the tamper-resistant SIM card of the subscriber. The symmetric K_i is generated by SIM manufacturer or by the MNO when the SIM card is programmed, so the operator also has a copy of this key. Three main types of cryptographic algorithms are used in 2G. These are the pair of algorithms denoted A3 and A8 (where the combined pair is commonly called COMP128) of which there exists four different sets, as well as the stream cipher algorithm A5. The key K_i is used with COMP128, or more specifically with the A3 algorithm for subscriber authentication, and with the A8 algorithm for generating the session cipher key CK. The key CK is then used with A5 for encrypting the data over the radio link between the base stations of the serving network (SN) and the handset denoted MS (Mobile Station).

The algorithm for encrypting the radio link in 2G is generally called A5, and there were originally two different versions called A5/1 and A5/2. The first algorithm A5/1 was developed in 1987 when GSM was not yet considered for use outside Europe, and second algorithm A5/2 was developed in 1989 specifically for markets outside Europe. The idea was that European countries should use the relatively strong (but know vulnerable) A5/1 algorithm, whereas markets outside Europe should use the weak A5/2 algorithm. Though both were initially kept secret, their general design was leaked in 1994 and the algorithms were entirely reverse engineered in 1999 from the firmware of a GSM telephone (Briceno, Goldberg, Wagner 1999).

The discriminating crypto policy of 2G reflected the mindset of the cold war and was accepted by mobile operators around the world, so network and phone manufacturers went ahead to start producing equipment that could use both A5/1 and A5/2. There is also the alternative of leaving the radio channel unencrypted, which is simply called A5/0. In addition, from around 2004 a third algorithm called Kasumi developed for UMTS 3G started to also be implemented in mobile phones and in 2G base stations where two of its variants are called A5/3 and A5/4.

When a phone connects to a base station an authentication and key agreement protocol (sequence of messages) is executed during call set-up. System entities belonging to the subscriber’s home operator are collectively called HE (Home Environment). System entities in the visited network are collectively called SN (Serving Network).

The SN sends the subscriber IMSI to the HE which looks up the profile of that specific subscriber in the HLR (Home Location Register) where it finds the secret individual subscriber key K_i . The HE computes a set of n cryptographic authentication vectors (AV) also called *GSM triplets* denoted $AV_{GSM} = \{CK, XRES, RAND\}$ consisting of the cipher key CK generated by the algorithm A3, the expected response XRES generated by the algorithm A8, and the random nonce RAND. The HE sends the set of AV_{GSM} vectors to SN which can use it for n authentication exchanges with the SIM. When the last AV_{GSM} has been used, a new set is requested from HE.

After receiving the AV_{GSM} vector, SN sends RAND across the radio link to the mobile phone denoted MS (Mobile Station) which in turn passes it to the embedded SIM chip. The SIM uses the algorithm A3 to compute the response RES, and the algorithm A8 to compute the cipher key CK, both as a function of RAND and the secret individual subscriber key K_i . RES is returned via MS across the radio link to SN which checks that $XRES = RES$ to authenticate the subscriber. The SIM also sends cipher key CK to MS. After successful subscriber

authentication the encrypted radio channel is established between SN and MS using key CK with one of the versions of the A5 algorithm, where the specific version of the algorithm is dictated by SN. This simple scenario illustrated in Figure 1.

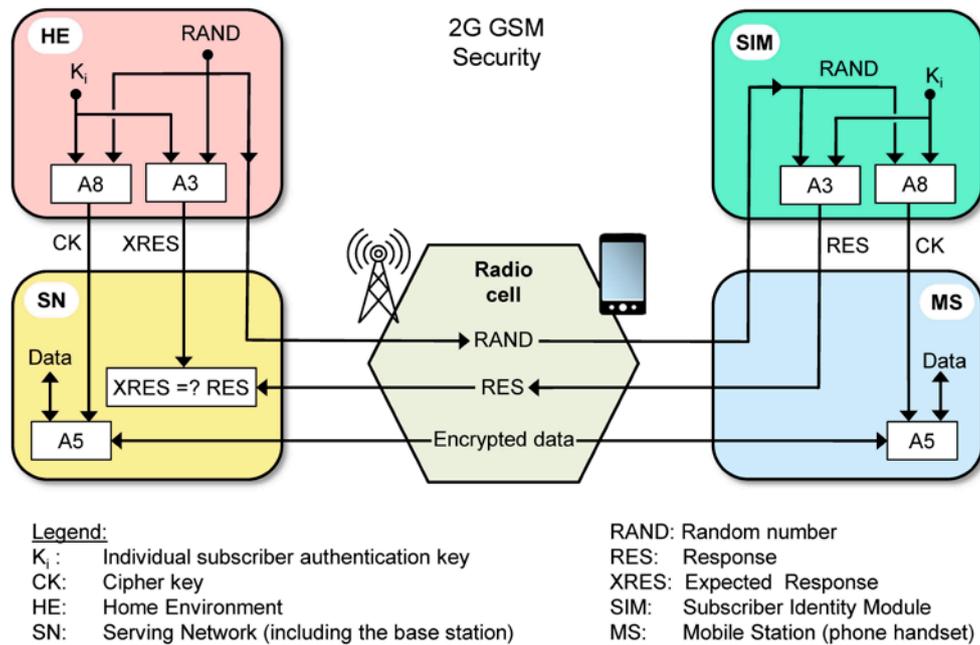


Figure 1. Security architecture in 2G GSM

The logistics of replacing the algorithms of COMP128 (A3 and A8) is relatively simple, and consists of distributing new SIM cards to subscribers of the MNO, and upgrading centralised system components in the MNO network. This can be done by one mobile network operator independently of other operators, and has typically been done several times by each operator. Unfortunately the situation is much worse for the A5 ciphers, which are embedded in the hardware of most handsets and in base station equipment around the world. Today it is possible to conduct practical attacks on A5/1 so that calls encrypted with A5/1 can easily be decrypted on a high-end PC. Due to the practical difficulty of replacing 7 billion mobile phone handsets, this vulnerability is very difficult to remove in the short to medium term. Moreover, this threat is aggravated by the fact that A5/1 was mandatory in every handset that supports GSM communication.

2.2. Security in 3G UMTS

3G UMTS security builds on elements of 2G by retaining the security features that worked well, by improving those that did not, and by adding new security features. As in GSM, a smart card called the USIM – which represents the subscriber – is inserted into the MS.

When SN detects a new MS in the network and receives the subscriber IMSI, an authentication data request is sent to HE which generates a set of n authentication vectors $AV_{UMTS} = \{RAND, XRES, CK, IK, AUTN\}$ consisting of a random number RAND, an expected response XRES, a cipher key CK, an integrity key IK and an authentication token AUTN. The array of n authentication vectors AV_{UMTS} is sent from HE to SN where it is stored in the VLR (Visited Location Register). Since an UMTS authentication vector consists of five components it is called a UMTS ‘quintuplet’ in analogy to the AV_{GSM} ‘triplet’ of 2G GSM.

In the UMTS AKA protocol (Authentication and key Agreement) SN first selects the next authentication vector from the array and sends the parameters RAND and AUTN to the USIM via MS. The USIM checks whether AUTN can be accepted by verifying that $MAC = XMAC$. The AUTN token can only be accepted if the sequence number contained in this token is fresh. This check can also be considered as an approval of the SN identity by the subscriber’s home operator, but falls short of being a proper network authentication by the USIM. After validating AUTN, the USIM returns response RES to SN. The USIM also computes CK and IK. The SN compares the received RES with XRES. If they match, the SN considers the authentication exchange as successfully completed. The USIM generated keys CK and IK are transferred from the USIM to MS, those

received by SN through AV_{UMTS} are sent from the SN VLR to the Base Station in the radio cell where the subscriber is located. These keys are then used by the ciphering and integrity functions in MS and in the SN Base Station as shown in Figure 2.

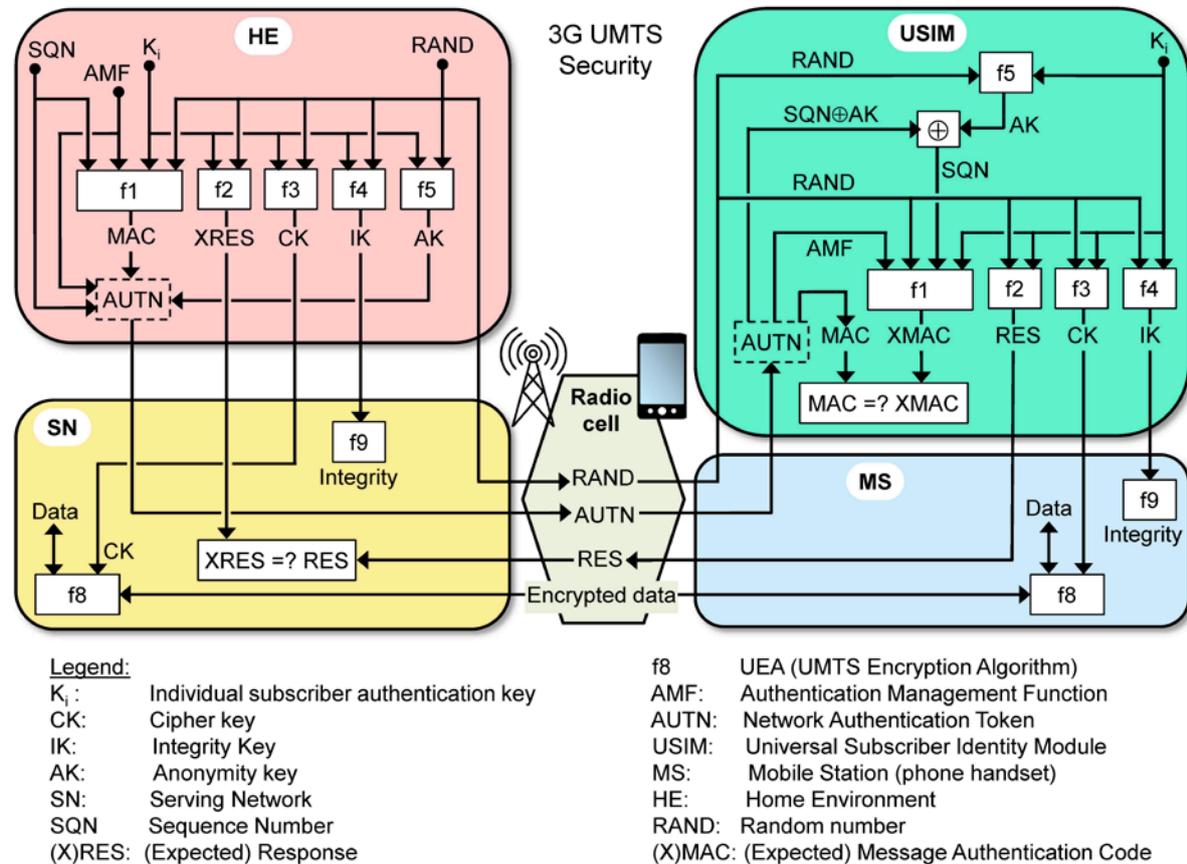


Figure 2. Security architecture in 3G UMTS

The algorithms and key derivation functions in UMTS have generic names denoted 'f#', where each function f# is implemented with a specific algorithm/function. The algorithms/functions f1 – f5 are located in the HE/USIM domain and can be chosen by each MNO independently of others, because no global alignment of these algorithms/functions is needed. In fact the UMTS standard does not dictate the specific algorithms/functions to be used for f1- f5, it only proposes some reference examples. This makes the global mobile network less vulnerable to specific cryptographic attacks because such attacks typically only affect a limited set of MNOs. In addition an MNO is able to replace any vulnerable algorithms with new and stronger ones.

However, the functions f8 and f9 have specific implementations that are dictated by the UMTS standard because of the requirement for global interoperability. As a result the exact same specific algorithms are implemented in every mobile handset worldwide. In particular f8 can be UEA1- the Kasumi algorithm which is also used in 2G GSM where it is called A5/3. Alternatively it can be UEA2 with is based on the newly introduced Snow 3G algorithm. It is also possible that SN and USIM negotiates to let f8 be instantiated as UEA0 which turns off encryption altogether, but this is normally only used for emergency calls. The specific algorithm used for f9 to provide integrity protection is either UIA1 (UMTS Integrity Algorithm 1) which is based on Kasumi, or UIA2 which is based on Snow 3G. Integrity can not be switched off, which means that one of these algorithms must be used.

Security in 3G UMTS is significantly stronger than in 2G GSM, but still has certain vulnerabilities. 3GPP has identified various potential threats categorised as DoS (Denial of Service), user impersonation, network impersonation, MitM (Man-in-the-Middle) attack, and identity catching (Mobarhan 2012). Of these threats, Man-in-the-Middle represent the most potent attack, which can be used to identify the IMSI with so-called IMSI catchers discussed in Section 3 below.

2.3. Security in 4G LTE

The 4G LTE architecture was developed by 3GPP taking into consideration security principles right from its inception and design based on five security feature groups (3GPP, 2011).

- (i) Network access security, to provide a secure access to the service by the user.
- (ii) Network domain security, to protect network elements and secure signalling and user data exchange.
- (iii) User domain security, to control the secure access to mobile stations
- (iv) Application domain security, to establish secure communications over the application layer
- (v) Visibility and configuration of security, allowing users to check if security features are in operation.

LTE is designed with strong cryptographic techniques, mutual authentication between LTE network elements with security mechanisms built into its architecture. Cryptographic protection is provided on many different layers in 4G, which requires a relatively large number of cryptographic keys. For that reason, a multi-level key hierarchy was introduced using multiple key derivation functions. The advanced security architecture puts higher requirement on security operations management by the MNOs (Mobile Network Operators). While the security in 2G GSM and 3G UMTS consists of a fixed set of standardised modules, in 4G LTE the MNO must to a large extent decide which security functionality it wants to implement, so that security management in fact becomes a challenge (Bhasker, 2013).

The authentication vector for LTE is a quadruplet denoted $AV_{EPS} = \{RAND, XRES, AUTN, K_{ASME}\}$ consisting of the random number (RAND), expected user response (XRES), authentication token (AUTN), and the Access Security Management Entity Key (K_{ASME}). The number of vectors provided by HE can be less than or equal to the number of AVs requested by the SN. In LTE 4G there is a proper key hierarchy based on the functions KDF, AKDF³ and BKDF³, and there are in fact many more keys than those shown in Figure 3.

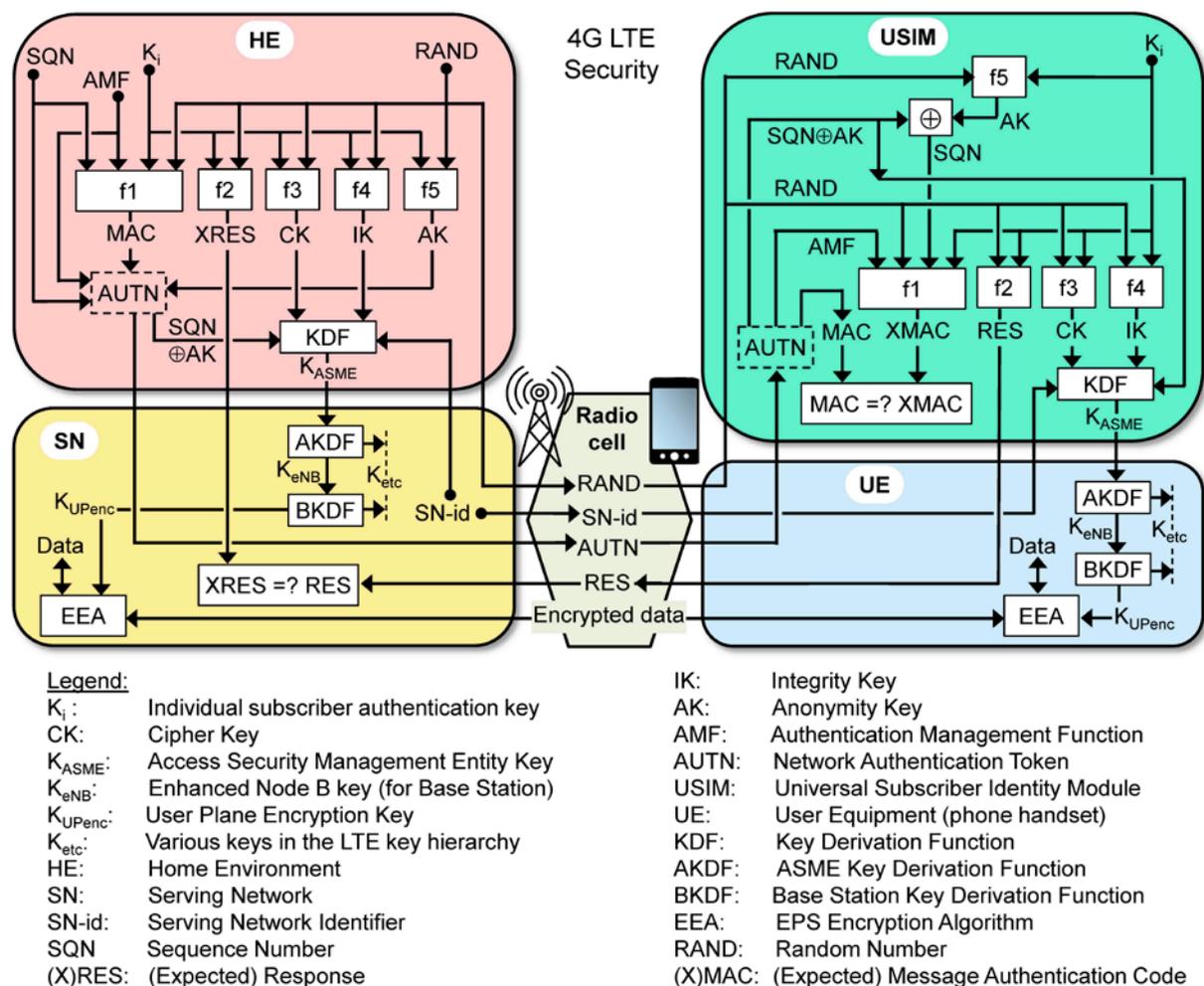


Figure 3. Security architecture in 4G LTE

³ Named so by us.

It can be seen that the key K_{ASME} depends on the network identity denoted SN-id which has been approved by the subscriber's home operator and used in the KDF function. This can be considered to be proper network authentication, in contrast to the method used in 3G UMTS which is only network approval, without necessarily knowing the network identity explicitly.

There are three different versions of the traffic encryption algorithm EEA (EPS⁴ Encryption Algorithm) denoted EEA1, EEA2 and EEA3, where the latter is used specifically in Chinese mobile networks. It is also possible to let EEA be instantiated as EEA0 which turns off encryption altogether, but this mode is only used for unauthenticated emergency calls. There are three versions of the integrity algorithm EIA (EPS Integrity Algorithm) denoted EIA1, EIA2 and EIA3 (not shown in Figure 3). Integrity can not be switched off, so one of the EIA algorithms must be used.

Cryptographic security in 4G LTE is considered by experts to be relatively strong, so that the most significant security vulnerabilities no longer are found in the architecture of Figure 3. Since 4G LTE introduces full IP capability many of the typical vulnerabilities of the Internet also become relevant for LTE. Potential threats of this category are described in (Macaulay, 2013), but are outside the scope of this paper.

3. Discussion

The mobile network is in fact an access network, either for accessing the PSTN (Public Switched Telephone Network) or for accessing the Internet. The purpose of accessing the PSTN is for making national or international voice calls, and the purpose of accessing the Internet is for accessing the vast resources available on the Internet. In order to understand the fundamental security problem underlying mobile networks it is useful to make a comparison with other access networks.

Wi-Fi is an access technology for accessing LANs (Local Area Networks) in an organisation as well as for accessing the global Internet. In that sense Wi-Fi is very analogous to mobile networks with regard to accessing the Internet. Wi-Fi security has evolved through several generations, similarly to mobile network security, where each new generation was developed in response to vulnerabilities found in the previous version. WEP (Wired Equivalent Privacy), introduced in 1999 was the first security technology for Wi-Fi. The intention of WEP was to provide data confidentiality comparable to that of a traditional wired network. However, serious security flaws were quickly discovered so that attackers could easily intercept or access other people's Wi-Fi access networks. In 2003 the Wi-Fi Alliance announced that WEP had been superseded by WPA (Wi-Fi Protected Access). In 2004, with the ratification of the full 802.11i standard (i.e. WPA2), the IEEE declared that WEP had been deprecated. What it meant was the WEP was no longer to be implemented and used in devices and Wi-Fi routers.

The contrast between mobile networks and Wi-Fi is pedagogic. For Wi-Fi access networks, previous generations of weak security technology was phased out after only 5 years, whereas in mobile access networks the first generation of weak security technology dating 25 years back is still in use. While modern and strong secure technology for mobile networks has been developed and is being used in the form of 3G and 4G, for various reasons the stakeholders in the mobile industry have decided to let weak security from 2G remain in the network. The consequence of keeping outdated security in the mobile network is that overall security is actually not strengthened by adding modern security technology because the outdated security technology represents a weakest link as illustrated in Figure 4.

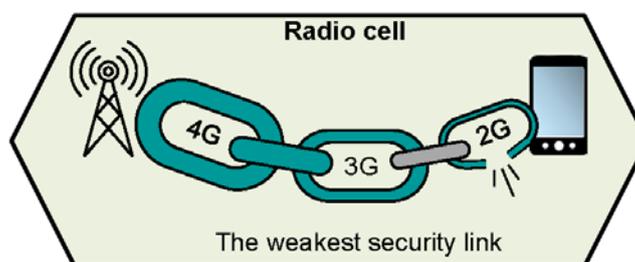


Figure 4. 2G GSM as the weakest security link in mobile networks.

In 2000, around 130 million GSM 2G customers relied on A5/1 to protect the confidentiality of their voice communications; and by 2011, it was 4 billion. In 2015 the number of mobile phones that can communicate over 2G is approximately equal to the world population of 7 billion.

⁴ EPS (Evolved Packet System) is the packet-based transmission and switching architecture developed for LTE.

When searching for reasons for why the weak security of 2G is still implemented in most mobile networks and in all mobile phones worldwide, the explanation seems to be a mix of business models and national security policies. The original policy of allowing weak or no encryption is still in force in the sense that many states around the world require the ability to request unencrypted radio traffic between mobile phones and the base stations, with the purpose of intercepting mobile phone traffic for law enforcement. Because the mobile networks share the same global standards, this has consequences for all other countries as well. In order to get optimal coverage a mobile phone must have 2G, 3G and 4G connectivity, and according to the standards must be able to send radio traffic in both encrypted form and in clear, where the network can decide which of these modes to use. If the phone could be configured to always encrypt, then it would be denied network access if it requests encryption in a country where encryption is not permitted. Manufacturers do not produce phones that only allow encrypted traffic or that do not support 2G GSM, because these phones would have inferior network coverage and therefore would not sell in the mass consumer market.

Security aware subscribers would be interested in knowing whether the radio traffic is encrypted or not, and the specifications for 2G, 3G and 4G published by ETSI and 3GPP actually do say that phones can give an alert to the subscriber in case of unencrypted traffic, where this alert should be triggered by the SIM/USIM. However, most MNOs de-activate this function in the SIM/USIM so that subscribers get no alert in case of unencrypted traffic (Paget, 2010). Disabling the alert function is understandable, as alerts would make many people confused or worried, and would most likely result in many people hanging up calls which thereby would lead to reduced revenue from network usage, and cause increased help desk calls which would be an extra burden for operators in terms of subscriber management.

When looking at Figure 1 it is obvious that MS (the mobile phone) actually knows whether the traffic is encrypted or not, as a function of which version of the 2G encryption algorithm is being used. In case MS uses A5/0 then the traffic is unencrypted, and in case MS uses A5/2 or A5/3 then the traffic is encrypted. A mobile phone could thus give alerts independently of triggers from the SIM, but no phone does. The rationale for phone manufacturers is similar to that of network operators. Many people would be confused and worried by a phone that gives alerts so they might be reluctant to use it and instead buy a phone from the competitor. The conclusion from this simple analysis is that neither MNOs nor phone manufacturers have any incentive for alerting subscribers in case of non-encrypted traffic.

The combination of national security policy and commercial business consideration is this that all mobile phones can send unencrypted traffic, and can be dictated to do so by the network operators, and that subscribers will not be alerted. This is in fact an ideal situation for IMSI catching and phone traffic interception.

The term 'IMSI catcher' denotes a fake base station which can be used to obtain the IMSI (International Mobile Subscriber Identity) from nearby mobile phones, and which in addition can intercept the radio traffic from the same mobile phones. An IMSI catcher pretends to be a 2G base station and sends out stronger signals than legitimate 2G base stations in the same area. As a result handset nearby determine the IMSI catcher to be the closest 2G base station and will send requests for connection. On first time connection the mobile phone must send the permanent IMSI, which is why the fake base station is called an IMSI catcher. On subsequent connection to the same fake base station, a pseudonymous TMSI is sent, but this does not help since the IMSI has already been caught. The IMSI catcher can dictate the settings for the connection, and is free to dictate the use of A5/0 which means unencrypted traffic, so that phone calls can be intercepted. The IMSI catcher must prevent phones from connecting to legitimate base stations with 3G or 4G, and can use radio jamming of the spectrum for 3G and 4G for that purpose. For mobile phones it is then as if no 3G or 4G network is available in the area, so they will connect to the IMSI catcher instead. In order to complete calls, the IMSI catcher must have a SIM and be able to connect to a legitimate base station nearby, so that it operates as a relay station between mobile phones and legitimate base stations.

IMSI catchers are normally only sold to national law enforcement organisations, but can easily be bought by individuals and private organisations. The price has dropped significantly in recent years. Originally they were sold for several hundred thousand dollars but can now be purchased for less than US\$1000. Most IMSI implementations are relatively bulky so that installation in cars is the most practical. However, body-worn IMSI catchers are also available (Goodin, 2013).

The policy of allowing interception of radio traffic for national law enforcement purposes necessarily has as consequence that criminal organisations and other nation states also are able to intercept mobile phone traffic. The balance of making security weak enough for national law enforcement organisations and at the same time strong enough to thwart attacks by criminal organisations is almost impossible to make. The rationale of having weak security must this be that the legal interception is more valuable than protecting subscribers against criminals and foreign intelligence organisations.

Assuming that 2G with its weak security will stay in mobile networks for years to come, it is worth considering mitigation strategies. Possible strategies are:

- 1) Use phones that can detect when the mobile network is being attacked.
- 2) Deploy sensors at strategic geographic places to detect fake base stations
- 3) Include technology and intelligence in every base station to detect fake base stations.

With regard to strategy (1) there are apps for various mobile phone operating systems available that can detect IMSI catchers. Alternatively, mobile phones can have this as an integrated function which can be activated by subscribers whenever needed.

Strategies (2) and (3) would require major expenditures by governments, individuals or MNOs, and it is a political question whether it should be implemented and how it should be financed. A citizen-oriented approach for (2) could be to use phones of strategy (1) to populate a crowd-based database of known fake base stations.

4. Conclusion

Our analysis has shown that policy and technology decisions regarding security made 25 years ago still determine the security level of mobile networks today. In other domains, such as Wi-Fi, old security technology is phased out and replaced with modern strong security technology. There is also modern strong security technology being introduced in mobile networks, but for business and political reasons it seems impossible to phase out the old insecure technology. The paradoxical consequence is that current mobile network traffic can be intercepted just as easily as it could 25 years ago.

We have also proposed a set of mitigation strategies for strengthening mobile network security. Security aware users can install apps for detecting attacks, mobile phone manufacturers can offer secure phones that detect and prevent attacks. Finally, governments and mobile network operators can deploy sensors in specific geographical areas for detecting attacks. In our opinion, a combination of these strategies should be encouraged.

References

- 3GPP (2011). 3rd Generation Partnership Project, "TS 33.401: System Architecture Evolution (SAE); Security architecture. Network, ver.11.2.0, release 11.," 3GPP, 2011.
- Bhasker, Daksha (2013). *4G LTE Security for Mobile Network Operators*. Journal of Cyber Security and Information Systems. Vol: 1 Num: 4 Published: October, 2013. Publisher: Cyber Security and Information Systems Information Analysis Center (CSIAC).
- Briceno, Marc, Goldberg, Ian and Wagner, David (1999). *A pedagogical implementation of A5/1*. <http://www.scard.org/gsm/a51.html>
- Goldberg, Ian and Wagner, David and Green, Lucky (1999). *The (Real-Time) Cryptanalysis of A5/2*. Rump session of Crypto'99, 1999.
- Goodin, Dan (2013). *The body-worn 'IMSI catcher' for all your covert phone snooping needs*. Ars Technica online. 1 September 2013.
- Macaulay, Tyson (2013). *The 7 Deadly Threats to 4G*. McAfee White Paper. McAfee, Inc. 2013.
- Mobarhan, Mojtaba Ayoubi and Mobarhan, Mostafa Ayoubi and Shahbahrami, Asadollah (2014). *Evaluation of Security Attacks on UMTS Authentication Mechanism*. International Journal of Network Security & Its Applications (IJNSA), Vol.4, No.4, July 2012.
- Paget, Chris (2010). Practical Cellphone Spying. DEFCON 18, Las Vegas, July/August 2010.