Security in Public Mobile Communication Networks

Hannes Federrath[1], Anja Jerichow[1], Dogan Kesdogan[2], Andreas Pfitzmann[1]

[1]*TU Dresden, Institut für Theoretische Informatik, 01062 Dresden*
[2]*RWTH Aachen, Informatik IV, Ahornstr. 55, 52074 Aachen*

Summary

Starting with the strongest data protection and security requirements for public radio networks known to us, the lack of data security in existing and planned public mobile telecommunication systems is shown. Possible solutions are demonstrated to show how such requirements can be met. After explaining principles representative for the research of the past we give further suggestions.

1 Introduction

The increasing use of mobile communication networks results in ever more stringent security requirements. In an information society, availability, integrity and confidentiality are essential. Especially the provision of the latter is hard to demonstrate. If someone or some component is able to collect and store personal data, one cannot be sure that this data is not gathered and not (mis)used. But this "being sure" is essential with respect to privacy and data protection. Therefore, legal means alone are insufficient and have to be complemented by technical means we are going to describe in the sequel.

This first part is introductory and reflects the structure of future mobile communication networks. When users communicate in such networks there is much data generated which needs to be handled in a secure way. We present data protection and security requirements which we believe are the strongest ever presented.

1.1 Structure of Future Mobile Telecommunication Systems

Telecommunication networks are of growing importance. Especially the world of mobile telecommunication systems is expanding very fast.

UMTS - the ***U**niversal **M**obile **T**elecommunication **S**ystem* - is very probable to be to the future the mobile communication network of the future in Europe. It is supposed to be a platform for existing systems of the second generation, e.g. standards like GSM, DCS1800, UPT, ERMES. Additionally it will allow integration of new developing systems of the third generation. At the same time worldwide standardization activities will lead to FPLMTS - the ***F**uture **P**ublic **L**and **M**obile **T**elecommunication **S**ystem.*

Services currently provided by single networks will be supported by a unified infrastructure. The design of UMTS is not yet fully defined. Already known is that UMTS will be based on the following three components: an access network, the fixed core network and an intelligent network (IN). The IN concept allows more flexibility. This is especially interesting for service providers. Because of the modular structure, definition and implementation of new services are possible in a very short time. Necessary mobility procedures such as handover and location update are realized by IN functions. B-ISDN is meant to be the underlying fixed core network. One advantage of using B-ISDN is the shared use of expensive network resources. Using satellite techniques for overlay-purposes is planned as well.

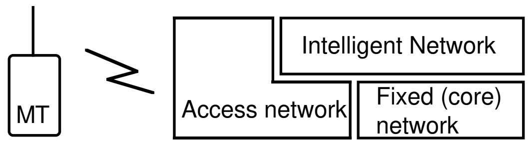

Figure 1 Structure of UMTS according to [1]

A stringent examination of existing security problems is necessary since public telecommunication networks and projects like UMTS and FPLMTS are developing fast. Solutions which solve or at least minimize these security problems are needed.

For this purpose existing knowledge of solutions in fixed communication networks can be used [2,3]. The differences to problems in mobile communication networks are the following:

- Bandwidth is and will be limited in radio networks since the electromagnetic spectrum can be used only once.
- It must be possible to find a moving mobile station.
- Not only user data and switching data (see also 1.2) which relate to a certain person are considered to be protected but also the location of a person who is using the mobile station.

Dividing the area into many cells such that the electromagnetic spectrum can be used more than once reduces the first problem. But at the same time it exacerbates the third problem. It is much easier to determine the exact location of a mobile station when searching in a small area of the cellular network.

1.2 User Data and Switching Data

Data to be protected when communicating can be subdivided in **user data** and **switching data.** User data is data given to the network to be transmitted. Switching data is data needed for connecting sender and recipient. Data of content, data of interest and traffic data can be derived from user and switching data.

In digital mobile systems the implementation of encryption is inexpensive because of digital data transmission and could therefore be easily applied. But, for example, systems of the GSM standard use only link-to-link encryption at the radio interface to protect user data. If encryption algorithms are not publicly known nobody can be sure that data cannot be read by third parties and/or data of interest can be filtered out. Obviously, the network operator has this possibility since he creates the encryption keys. For example, in GSM switching data is stored in registers (*home location register* HLR, *visitor location register* VLR). This data has to be transmitted for billing. This data exchange cannot easily be traced especially if network operators and service providers are not the same institution.

If such data is accumulated it allows conclusions about the interests of network users. It can tell who has communicated with whom how long. Traffic data allow to construct location profiles.

1.3 Requirements resulting from Data Protection

The *ability* to collect and analyze data of others on a large scale, we consider as a violation of personal privacy – since there are possibilities to design and implement communication systems in such a way that they do not enable this.

For public mobile communication systems intended for broad use, in our opinion, the following requirements resulting from data protection should be met:

Protection of Confidentiality

- c1 *Message contents* should be kept confidential towards all parties except the communication partners.
- c2 *Sender* and/or *addressee* of messages should stay anonymous to each other, and third parties (including the network operator(s)) should be unable to observe their communication.
- c3 Neither potential communication partners nor third parties (including the network operator(s)) should be able to locate mobile stations or their users.

Protection of Integrity

- i1 Forging *message contents* (including *sender's address*) should be detected.
- i2 The recipient of a message *y* should be able to prove to third parties that entity *x* has sent message *y*.
- i3 The sender of a message should be able to prove the sending of a message with correct contents, if possible, even that the addressee received the message.
- i4 Nobody can cheat the network operator(s) with in terms of usage fees. But on the other hand, the network operator(s) can only demand usage fees for correctly delivered services.

Protection of Availability

- a1 The communication network enables communication between all parties who wish to communicate (and who are allowed to).

Such data protection requirements cannot be fulfilled by legal means alone – and laws, which cannot be enforced, have a negative effect of law-abiding in general. Confidentiality requirements must therefore be enforced by the **prevention of the gathering of personal data**. There is no other way known to achieve privacy with respect to the operator and designer of network components.

With respect to the latter, it is mostly completely ignored that a component or system might be under control of everybody who has had access to it so far, because he might have installed a Trojan Horse. Not only the designer but also every complex tool used to design the system might be able to do so. Moreover, the Trojan Horse in the first tool may be implemented by another tool used to generate the first one and so on (transitive Trojan Horse).

In the following, we will show how by technical means for data protection we cannot only provide security for the network operator, but for the users of the network as well.

2 Realization of Data Protection Requirements

How and where is it possible to realize data protection requirements? The fact of mobility makes it difficult to apply well known concepts in the same way as in fixed networks. After outlining these basic concepts some ideas are given which could point into the right direction.

2.1 Basic Concepts

Security problems may be solved by using methods such as end-to-end-encryption and link-to-link-encryption. The anonymity of participants can be protected by using certain kinds of addressing, broadcast and other methods for example MIXes and superposed sending.

2.1.1 Protection of User Data

Requirement c1 means, trusted communication between two participants of the same and of other networks must be possible. The same must be true for integrity requirements i1, i2 and i3. These can be achieved by encryption, digital signatures and authentication codes. In fact, c1 can be accomplished by end-to-end-encryption. For i1 to be realized, for example, a hash-value of a message is digitally signed. For i3 a digital signature of the sender of the message is necessary. Fulfillment of requirement i3 needs a signed receipt from the recipient or the message transmission system.
Cryptography is only applicable if the following conditions are true:

The different services and different network systems need to match corresponding encryption methods and protocols. But this seems to be more a legal (political) and economical than a technical problem.

User channels (according to the OSI-7-layer-model of the ISO at the transport layer 4 or higher) must be bit-transparent, i.e. bits to be transmitted on the signal path must not be changed or interfered with. The minor change of bits could mean a loss of integrity on the signal path. Furthermore, a change of only one bit would be followed by an increased rate of errors since encryption systems usually produce a strong dependency between bits.

Even bit-transparency is not implemented in every already realized and standardized network, e.g. systems of the GSM standard have non bit-transparent speech channels but in network systems like ISDN bit-transparent channels are available. Considering these aspects the integration of networks and services must be planned carefully.

2.1.2 Protection of Switching Data

The following concepts show the possibility of developing networks which fulfill our data protection requirements.

2.1.2.1 Link-to-Link Encryption

The contents of a message can be sufficiently hidden by end-to-end encryption at the ISO layer 4. If protocols of the layers 1 to 3 also contain personal data then it is also necessary to protect this information by link-to-link encryption. This information could be the address of a mobile terminal or the address of a smart card. This data is strongly related to the owner because such equipment is of a very personal character and will not be changed after every usage.

2.1.2.2 Recipient Anonymity by Broadcast and Addressing Attributes

Receiving a message can be made completely anonymous to the network by delivering the message (possibly end-to-end-encrypted) to all stations (broadcast). If the message has an intended recipient, a so called addressee, it has to contain an attribute by which he and nobody else can recognize it as addressed to him [4]. This attribute is called an *implicit address*. It is meaningless and only understandable by the recipient who can determine whether he is the intended person. In contrast, an *explicit address* describes either a place in the network or the place of a station.

Implicit addresses can be distinguished according to their visibility, i.e. whether they can be tested for equality or not. An implicit address is called **invisible**, if it is only visible to its addressee and is called **visible** otherwise.

		address distribution	
		public address	private address
implicit address	invisible	very costly, but necessary to establish contact	costly
	visible	not advisable	frequent change after use

Figure 2 Combination of implicit addressing modes and address distribution

Invisible implicit addresses, unfortunately very costly, can be realized with a public key cryptosystem. *Visible implicit addresses* can be realized much easier: Users choose arbitrary names for themselves, which can then be prefixed to messages.

Another criterion to distinguish implicit addresses is their distribution. An implicit address is called *public*, if it is known to every user (like telephone numbers today) and *private* if the sender received it secretly from the addressee either outside the network or as a return address or by a generating algorithm the sender and the addressee agreed upon [4,5].

Public addresses should not be realized by visible implicit addresses to avoid the linkability of the visible public address of a message and the addressed user.

Private addresses can be realized by visible addresses but then each of them should be used only once.

2.1.2.3 Sender Anonymity by Using DC-Network or MIX-Network

A powerful scheme for sender anonymity is **superposed sending** which is published in [6,7] and is called **DC-network** (dining cryptographers network) there. For DC-networks it is proved that the anonymity of the sender is protected as long as station are linked by exchanged keys unknown to the attacker.

Unlinkability of sender and recipient can be realized by a special network station, a so called **MIX**, which collects a number of messages of equal length from many distinct senders, discards repeats, changes their encodings, and forwards the messages to the recipients in a different order [8]. This measure hides the relation between sender and recipient of a message from everybody but the MIX and the sender of the message. Change of encoding of a message can be implemented using a public-key cryptosystem. Since decryption is a deterministic operation, repeats of messages have to be discarded. Otherwise, the change of encoding does not prevent tracing messages through MIXes: Simply count the number of copies of each message before and after the MIX.

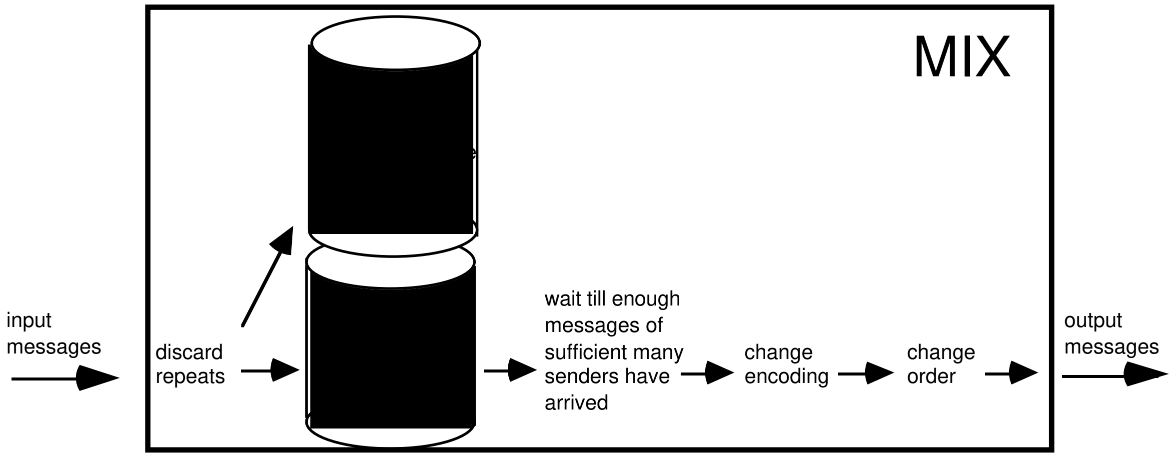

Figure 3 Functionality of MIXes

By using more than one MIX to forward a message from the sender to the recipient, the relation is hidden from all attackers in the network who do not control all MIXes which the message passed, nor have the cooperation of the sender [8]. MIXes should be independently designed and produced and should have independent operators, otherwise a single party is able to control a communication.

One method of achieving sender anonymity, i.e. of making unclear when a message was sent, is **dummy-traffic.** That means each user station sends among the real messages a number of meaningless messages.

Unfortunately the above described concepts are hardly feasible for the protection of the radio interface, e.g. dummy-traffic is not acceptable due to limited accumulator capacity and bandwidth. But, most of the concepts to protect switching data and interest data can be realized at the part of a communication system which is the fixed network (see 2.2.2).

2.1.3 Possibilities for Anonymous, Secure Accounting

In a public network there should exist a convenient and secure way to handle fees for used services (i4). Accounting mechanisms should be organized in a manner that anonymity and unobservability in communication networks are guaranteed (c2, c3).

In principle, there are two possible approaches: *individual accounting* and *flat-rate accounting*. The latter can be handled freely, i.e. without taking care of anonymity because no interesting subscriber-specific information is needed.

By using **flat-rate accounting** problems of fraud will be avoided. Furthermore, there is no need for a complex accounting management system. If there is enough bandwidth, as this is the case in fixed networks, flat-rate accounting can be used for local calls.

Individual accounting means to apply methods which preserve the anonymity of the person. It may be accomplished by installing an **unmanipulatable accounting system** placed at a fixed location combined with a mobile part. For example the fixed part can be placed at home and the mobile part in a mobile unit known as SIM (Subscribers Identity Module). Accounting can also be done by digital paying on-line towards the fixed part or by first buying a smart card like a pre-paid telephone card, that accumulated a number of service to be used. Main advantage of an unmanipulatable accounting system is, that the network does not need to manage accounting. Otherwise, manipulation at the fixed accounting equipment and at the mobile equipment could be done by the subscriber. These manipulation or a change of the accumulated number on the SIM must be recognized by network operator. Another approach of individual accounting is the use of an **anonymous digital accounting system** [14]. This guarantees anonymity and unobservability. It is important to design the protocol outlines in a way that nobody can use the system in a fraudulent way because anonymous digital accounting systems prevent prosecution after fraud took place.

2.2 Further Suggestions for Security in Digital Mobile Networks

Factors such as limited bandwidth on the radio interface, the ability of locating radio waves, registration of location information and working on small, low-power mobile equipment need new ideas for data security.

2.2.1 Mobile Radio Systems with Reduced Locating-Ability

In radio networks electromagnetic waves are the carrier of information. The source of the wave corresponds to the location of the mobile station. With locating methods it is possible to trace a mobile subscriber and record a moving-profile (or moving-track).

For defending (against this attack) we use the following model: A problem on processing electromagnetic waves are interferences, jam and noise. Substantial portions of noise are distributed with an equal power-density in a broad spectrum.

To locate the source of a wave it has to be detected. That means, the signal-to-noise-ratio must be above a definite level (value). This leads us to spread spectrum (SS) systems. SS base on the principle of information theory: The representation of a bit is not important for transmission. However, the energy (or power) of the spectral area is!

With a proper modulation the signal power density can be reduced to a level lower than the noise power density, i.e. the signal power density will be distributed to a broad bandwidth. A special SS is direct-sequence SS (DS/SS). First the user data is modulated to a narrowband carrier in a conventional way. In a second modulation the emerged signal is multiplied by a binary broadband pseudorandom-sequence, called pseudonoise (PN)-code. The PN-code is derived from a PN-key with a PN-code-generator. The PN-key is a secret attribute of the mobile transceiver (sender) and the intended recipient (receiver). The second modulation step produces a broadband signal with low power-density.

The intended recipient derives the PN-code from the secret PN-key, too. He multiplies the received signal by the PC-code and thus obtains the narrowband carrier signal back.

If orthogonal PN-codes are used for spreading more than one user can co-exist in the same frequency band with conventional narrowband signals. Thus, the spectral efficiency is equal to conventional time division multiplex or frequency multiplex.

From all this follows: Just as the signal power density is less than the noise power density, i.e. the signal-to-noise-ratio is small, so DS/SS signals are not detectable and not locatable if the PN-code is secret. Simply a *radiometer* can detect a radio wave by integration of noise in a spectrum over a longer time. But in this manner detected signals are not locatable. For further information see [9] and [10].

2.2.2 Trustworthy Maintenance of Location Information

For reachability of mobile subscribers in cellular radio networks (GSM, DCS-1800) location information has to be maintained by the net, usually in a *home location register* and *visitor location register*. Therefore, the network operator is able to record moving-tracks of users. That offends against our requirement c3.

Suppose that location information is maintained in a trustworthy environment and the network operator gets this information only for call-setup, then the problem is decreased.

Suppose the complete location management (e.g. location update, handover) is processed in a trustworthy environment and the linkability between sender and recipient is prevented (e.g. by MIXes) recording of moving-tracks would be impossible (or possible with high expense/time).

Such a trustworthy environment could be the fixed station (such as the conventional telephone) of a mobile subscriber in the fixed public telephone network. This station needs a sufficient capacity for encryptions, computations and management operations. We call this fixed station the **home personal computer HPC**. This HPC need not located to the home of the mobile subscriber. Just a trusted organization, company, institution can undertake the tasks of a HPC. Also it would be possible to generate two chips in cooperation. One chip for the subscriber corresponds to the *subscriber identity module* (SIM) in current used mobile networks. The other chip for the (mobile) network operator contains the functionality of the HPC.

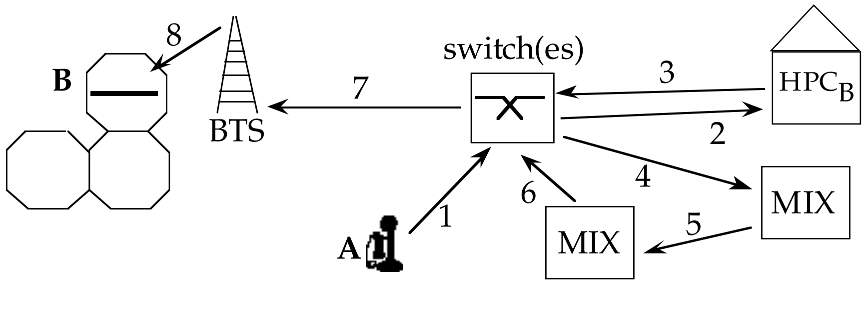

Figure 4 Trustworthy maintenance of location information in a home personal computer (HPC) and MIXes for prevention of linkability between sender and recipient

For example, **A** (see figure 4) establishes a connection to a mobile subscriber **B**. He is connected to the home personal computer of **B** (HPC_B) by switches (and possibly by MIXes). HPC_B routes the call to the base transceiver station BTS by using MIXes. Why do we prevent the link between HPC_B and BTS by MIXes? Otherwise, the network operator is able to analyze the traffic and can find out that the two parties communicate. Thus, he knows the subscriber is located in the supply area, i.e. our requirement c3 is violated.

2.2.3 Using Hierarchical Multi Layered Cell Architecture for Signaling

From our point of view the integration of different systems with different cell architecture is also a good possibility for realizing data protection. By giving an example, we want to demonstrate this new opportunity.

Currently, to set up an incoming call from the public switched telephone network to a mobile subscriber, the call must be broadcasted in the location area of the subscriber. If there is any precise routing information, the call can be forwarded directly towards a special BTS (base transceiver station), where the broadcast area is very small. Management of the subscribers location update is done by a protocol. This needs some signaling effort while the subscriber is roaming. In our days cell areas become smaller and smaller because bandwidth of the air interface is limited.

Towards the realization of UMTS the integration of different cellular networks with their different cell radii becomes unavoidable. Likely radii are Picocells (<100 m), Microcells (<1 km) and Macrocells (<35 km) [11]. The low altitude satellite network (LEO-satellite) appears as a complementary system to the different mobile cell networks. For this reason, Motorolas well known Project IRIDIUM will be put into service in 1998 with 66 satellites [12]. IRIDUM satellites will cover the whole earth surface. Giving an example for the broadcast area of a satellite system, the coverage area radius of ESA-developed medium altitude global satellite system (MAGSS-14) is considered to be 930 km.

A wide broadcast area of the overlaid network implies less need of exact location information which results in more anonymity. For secure networks that will lead to the following rule: an incoming call to the mobile subscriber should always be broadcasted in the widest broadcast area. We call this the "**wide area to small area rule**".

Figure 5 illustrates the above rule and depicts the message flow required to route an incoming call from the public switched telephone network **A** to a mobile subscriber **B**. To handle the incoming call to **B** it would be switched to a satellite gateway and finally broadcasted from a LEO-satellite. Because of the wide coverage area of the satellite, no exact knowledge of the present location of the mobile subscriber is needed.

After successfully contacting **B**, this subscriber will react appropriately and establish communication. Of course the subscriber **B** wants to establish communication in a way that is most favourable for him. Most subscribers would like the cheapest connection, because they must pay for the service. For this and other considerations like locally circumstances, e.g. the fact that terrestrial cellular networks are not operating in every place on earth, the subscriber chooses the next best linkage to a network. Usually, as shown in the figure, the mobile subscriber is interested in the next best available cellular network with the smallest cell radius. We call this: "**small area to wide area rule**".

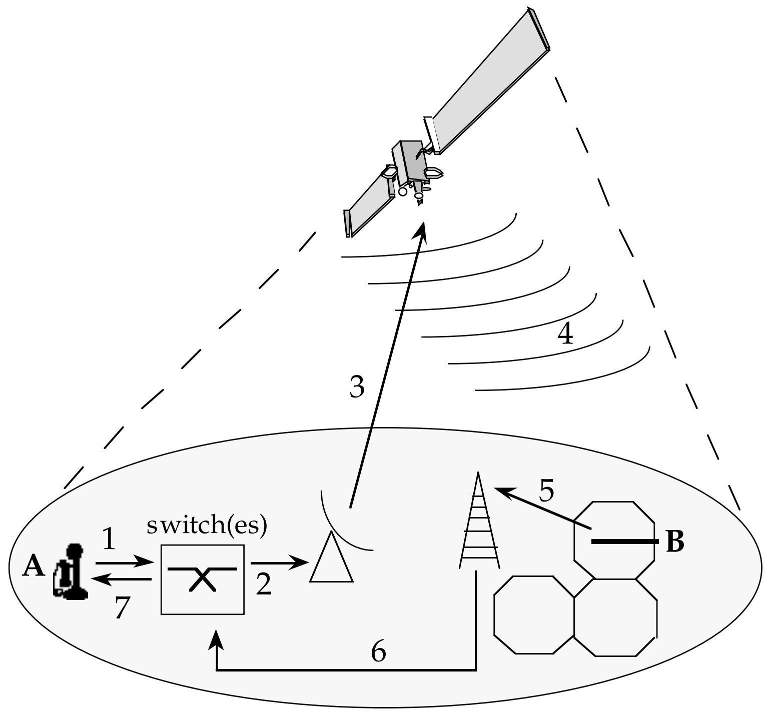

Figure 5 Call-setup and connecting to a mobile station

For the sake of simplicity we have chosen a star topology in the figure. A fixed satellite with one coverage/service area (oval area) is shown. By casting a message to a great number of subscribers, a perfect anonymity in the sense of information theory is given [13]. An eavesdropper could not locate a subscriber reacting to a call. But an active attack would destroy the broadcasted signal, and the availability requirement is at risk. This risk can be reduced if spread spectrum is used, because of the robustness of spread spectrum against distortion.

Broadcasting the signaling message can be performed by LEO-satellites in the same way, but with the restriction of more risks of information security aspects and some technical problems. Due to the high range of velocity of satellites (19˙000 to 25˙000 km/h), the visibility of a satellite at a fixed point on earth is less then 3 minutes. To continuously serve a coverage area, a mapping of the broadcast area to the corresponding LEO-satellites is needed. An active attack by the operator can threat the consistency of broadcasting. For example, the operator can send a signal via several satellites to a certain subscriber by broadcasting in a smaller broadcast area than normally covered. If the subscriber reacts then he is in this area.

A suitable countermeasure for this problem is to order the satellites to make a time stamp at each message with a digital signature.

2.2.4 Reachability Management

The new mobile communication systems do provide means to fulfill the need for personal mobility and reachability. This leads to new requirements in the field of reachability management. It would be desirable to develop a new concept and architecture for an advanced reachability manager. In their different roles such as in their spare time or at work, mobile subscribers must have the chance to filter communication demands. For example, a private digital assistant could handle these communication demands and work as an interface between protected subscribers and network.

The reachability manager could also contain the private data of subscribers (present location area) and data giving information about the relationship of a subscriber to other users. One realization could be the protection in a HPC (see 2.2.2).

3 Conclusions

This exposition gives options on how to design public communication networks to technically secure data protection. Furthermore, due to new requirements concerning personal mobility and reachability, the known techniques must be reconsidered and improved. Additionally, new techniques must be developed.

In the phase of standardization of the next generation of mobile networks, there is a chance to include these techniques as a solid part of the system. Otherwise, an unnecessarily great effort will be needed to implement data protection afterwards.

To protect the subscribers location in a cell, spread spectrum must be applied. Using decentralized organization of mobility management make it unnecessary to keep registers for managing location information. Introducing fixed home stations of users for reachability and location management can easily be performed while the next integrating step of communication technology is realized.

We are pleased to thank the German Science Foundation (DFG) and the Gottlieb Daimler- and Karl Benz-Foundation for financial support. Many thanks for all comments to Raschid Karabek, Sven Martin, Alexander Stahl and Frank Zündorff.

4 References

[1] Mitts, Hakan: Universal Mobile Teecommunication System – Mobile access to Broadband ISDN; Broadband Islands '94: Connecting with the End-User; W. Bauerfeld, O. Spaniol, F. Williams (Editors), Elsevier Science B.V. 1994, pp. 203-209

[2] Pfitzmann, Andreas; Waidner, Michael: Networks without user observability; Computers & Security 6/2 (1987), pp. 158-166

[3] Pfitzmann, Andreas; Pfitzmann, Birgit; Waidner, Michael: ISDN-MIXes – Untraceable Communication with Very Small Bandwidth Overhead; Proc. IFIP/Sec'91, May 1991, Brighton, North-Holland, Amsterdam 1991, pp. 245-258.

[4] Farber, D. J.; Larson, K. C.: Network Security Via Dynamic Process Renaming; Fourth Data Communications Symp., Oct. 1975, Quebec City, Canada, pp. 8-13 – 8-18

[5] Karger, P. A.: Non-Discretionary Access Control for Decentralized Computing Systems; Master Thesis, MIT, Laboratory for Computer Science, May 1977, Report MIT/LCS/TR-179

[6] Chaum, David: The Dining Cryptographers Problem: Unconditional Sender and Recipient Untraceability; Journal of Cryptology Vol. 1, Nr. 1, 1988, pp. 65-75

[7] Chaum, David: Security Without Identification: Transaction Systems to Make Big Brother Obsolete; CACM Vol. 28, Nu. 10, Oct. 1985, pp. 1030-1044

[8] Chaum, David: Untraceable Electronic Mail, Return Addresses, and Digital Pseudonyms; Communications of the ACM 24/2 (1981) 84-88

[9] Torrieri, D.J.: Principles of Secure Communication Systems (Second Edition); Artech Hoste, Boston, London, 1992

[10] Pickholtz, R.L.; Schilling, D.L.; Milstein, L.B.: Theory of Spread-Spectrum-Communications – A Tutorial. IEEE Transactions on Communications (1982) vol. 30, No.5, pp. 855-878

[11] Evci, C. Cengiz : Race-UMTS for third generation wireless communications, ANN. TÉLÉCOMMUN., 47, Nr. 7-8, 1992

[12] Flohr, Udo: Trabanten im All, iX 12/1994

[13] Pfitzmann, Andreas: Diensteintegrierende Kommunikationsnetze mit teilnehmerüberprüfbarem Datenschutz; Universität Karlsruhe, Fakultät für Informatik, Dissertation, Feb. 1989, IFB 234, Springer-Verlag, Heidelberg 1990

[14] Bürk, Holger; Pfitzmann, Andreas: Value Exchange Systems Enabling Security and Unobservability; Computers & Security 9/8 (1990) 715-721.