Chapter 7

Authentication codes

7.1 Introduction

The protection of unauthorized access to sensitive information has been a prime concern throughout the centuries. Still, it was not until Shannons work in the late 40's a theoretical model for secrecy was developed. Shannon's work was based on the concept of *unconditional security*, by which we mean that the faced enemy is assumed to have access to infinite computing power. Under this assumption Shannon developed some rather pessimistic results on the requirements for a cryptosystem to be secure.

More recently, we have understood that one usually needs to protect data not only against unauthorized access but also against unauthorized modifications. In a communication situation, we need to *authenticate* our transmitted messages. We need to check that they are indeed sent by the claimed sender and that they have not been modified during transmission. The threat from the enemy can be viewed as "intelligent noise", the noise taking the worst possible value for the sender and receiver. This means that error correcting codes will not help (because the noise just changes a transmitted codeword to another codeword), but we must introduce secret *keys* that are known to the sender/receiver but unknown to the enemy.

Authentication of transmitted messages can be done in (at least) three fundamentally different ways. We refer to them as *unconditionally secure authentication codes*, *message authentication codes*, and *digital signatures*.

Unconditionally secure authentication codes is the only solution to the authentication problem for an enemy with unlimited computing power. As this is the topic of the article, we continue the discussion in the next section.

Message authentication codes (MACs) refer to authentication techniques that use symmetric cryptographic primitives, i.e. block ciphers and hash functions, to provide authentication. As for unconditionally secure authentication codes, the sender and receiver are here assumed to share a common secret key. MACs is a very common authentication technique in for example banking transactions. MACs appear in many standards, and some common modes of operations for block ciphers provide MACs. Comparing with unconditionally secure authentication codes, MACs are not secure against an unlimited enemy. But they have other practical advantages, such as being able to authenticate many messages without changing the key.

Finally, digital signatures is an asymmetric solution. This means that the sender has a secret signing key of his own and the receiver has access to a corresponding public verification key. The sender first hashes the message to be transmitted using a cryptographic hash function. The result is then signed by a signature scheme. Common hash functions are MD5, SHA-1, etc., and common signature schemes are RSA and DSA. Digital signatures possess several advantages compared to the other two authentication techniques. Since it is an asymmetric technique, there is no need to distribute or establish a common secret key between the sender and the receiver. Basically, the sender generates his secret signing key and the corresponding public verification key. The verification key can then be presented in public. This means that anyone can verify the authenticity of a message. This leads to the second important difference, referred to as *nonrepudiation*. Since the sender is the only person able to generate an authentic message (the receiver cannot), we know that if a message is authentic it must have been generated by the sender. If the receiver has received an authentic message, the sender cannot deny having sent it. This somewhat resembles a handwritten signature, once you have signed you cannot later deny having signed. There are also drawbacks. Signature schemes rely on the hardness of problems like factoring and taking discrete logarithms. This means that we must work with very large numbers, which make the solutions slow compared to the other techniques, especially for short messages.

7.2 Authentication Codes

An unconditionally secure solution to the authentication problem first appeared in 1974 when Gilbert, MacWilliams and Sloane published their landmark paper "Codes which detect deception". As mentioned in that paper, Simmons was independently working with the same problems. In the beginning of the 80's Simmons published several papers on the subject which established the authentication model. Simmons work on authentication theory has a similar role as Shannons work on secrecy.

This section deals with unconditionally secure authentication codes. We provide some fundamental definition and results. We also include some common constructions. We start by presenting the mathematical model of unconditionally secure authentication due to Simmons.

The communication model for authentication includes three participants, *the transmitter*, *the receiver*, and *the opponent*. The transmission from the transmitter to the receiver takes place over an insecure channel. The opponent, who is the enemy, has access to the channel in the sense that he can insert a message into the channel, or alternatively, observe a transmitted message and then replace it with another message. The authentication model is illustrated in Figure 7.1.

The information that the transmitter wants to send is called a *source message*, denoted by s and taken from the finite set $\mathcal{S}$ of possible source messages. The source message is mapped into a (channel) *message*, denoted by m and taken from the set $\mathcal{M}$ of possible messages. Exactly how this mapping is performed is determined by the secret *key*, which is denoted by e and taken from the set $\mathcal{E}$ of possible encoding rules. The key is secretly shared between the transmitter and the receiver.

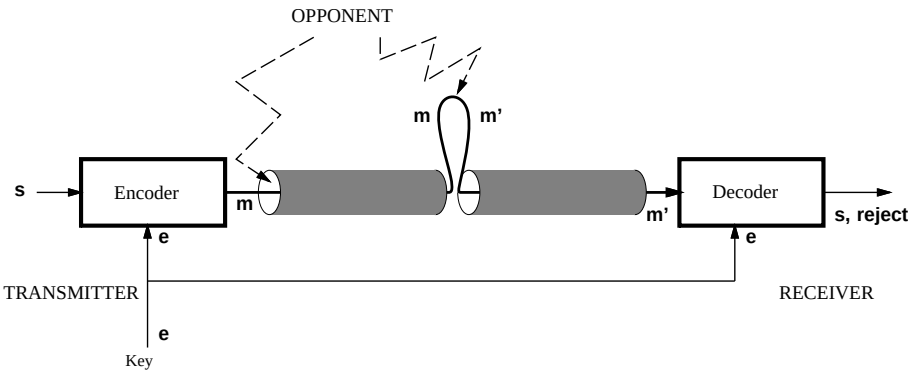

Figure 7.1: The authentication model.

Each key determines a mapping from $\mathcal{S}$ to $\mathcal{M}$. Equivalently, the encoding process can be described by the mapping f, where

$$f : \mathcal{S} \times \mathcal{E} \to \mathcal{M}, \quad (s, e) \mapsto m. \tag{7.1}$$

An important property of f is that if $f(s, e) = m$ and $f(s', e) = m$, then $s = s'$ (injective for each $e \in \mathcal{E}$). Two different source messages cannot map to the same message for a given encoding rule, since then the receiver would not be able to determine which source message was transmitted. The mapping f together with the sets $\mathcal{S}$, $\mathcal{M}$ and $\mathcal{E}$ define an *authentication code* (A-code).

When the receiver receives a message m, he must check whether a source message s exists, such that $f(s, e) = m$. If such an s exists, the message m is accepted as authentic (m is called valid). Otherwise, m is not authentic and thus rejected. We can assume that the receiver checks $f(s, e)$ for all $s \in \mathcal{S}$, and if he finds $s \in \mathcal{S}$ such that $f(s, e) = m$ he outputs s and otherwise he outputs a reject signal.

The opponent has two possible attacks at his disposal, the *impersonation attack* and the *substitution attack*. The impersonation attack simply means inserting a message m and hoping for it to be accepted as authentic. In the substitution attack the opponent observes the message m and replaces this with another message m', $m \neq m'$, hoping for m' to be valid.

We assume that the opponent chooses the message that maximizes his chances of success when performing an attack. The probability of success in each attack is denoted by P_I and P_S, respectively. They are more formally defined by [1]

$$P_I = \max_m P(m \text{ is valid}) \tag{7.2}$$

and

$$P_S = \max_{\substack{m,m' \\ m \neq m'}} P(m' \text{ is valid} | m \text{ is valid}). \tag{7.3}$$

Note that this definition considers only transmission of a single message. For transmission of multiple messages, we must introduce a more general definition of the deception probabilities.

Continuing, we define the *probability of deception* P_D as $P_D = \max(P_I, P_S)$. It is convenient to define $\mathcal{E}(m)$ as the set of keys for which a message m is valid,

$$\mathcal{E}(m) = \{e \in \mathcal{E}; \exists s \in \mathcal{S}, f(s, e) = m\}. \tag{7.4}$$

Let us now derive some basic properties for authentication codes. We see that of all the messages in $\mathcal{M}$, at least $|\mathcal{S}|$ must be authentic, since every source message maps to a different message in $\mathcal{M}$. Similarly for the substitution attack, after the observation of one legal message, at least $|\mathcal{S}| - 1$ of the remaining $|\mathcal{M}| - 1$ messages must be authentic. Thus we have two obvious bounds.

[1] We abbreviate expressions like $\max_{m \in \mathcal{M}}$ as $\max_m$ when no confusion occurs.

Theorem 7.1. *For any authentication code,*

$$P_I \geq \frac{|\mathcal{S}|}{|\mathcal{M}|}, \tag{7.5}$$

$$P_S \geq \frac{|\mathcal{S}|-1}{|\mathcal{M}|-1}. \tag{7.6}$$

From Theorem 7.1 we observe two fundamental properties of authentication codes. Firstly, in order to have good protection $|\mathcal{M}|$ must be chosen much larger than $|\mathcal{S}|$. This affects the message expansion of our authentication code. For a fixed source message space, an increase in the authentication protection implies an increased message expansion. The second property is that a complete protection, i.e., $P_D = 0$, is not possible. We must be satisfied with a protection where P_D is small.

Example: Let an authentication code with $\mathcal{S} = \{H, T\}$, $\mathcal{M} = \{1, 2, 3, 4\}$ and $\mathcal{E} = \{0, 1, 2, 3\}$ be described by the following table

		m			
	s	1	2	3	4
	0	H	T	-	-
e	1	T	-	H	-
	2	-	H	-	T
	3	-	-	T	H

.

It is easy to verify that $P_I = P_S = 1/2$ if the keys are uniformly distributed.

We assume that the reader is familiar with the basic concepts of information theory. As usual, $H(X)$ denotes the entropy of the random variable X, and $I(X;Y)$ denotes the mutual information between X and Y. We are now ready to state the next fundamental result in authentication theory, namely Simmons' bounds.

Theorem 7.2 (Simmons' bounds)**.** *For any authentication code,*

$$P_I \geq 2^{-I(M;E)}, \tag{7.7}$$

$$P_S \geq 2^{-H(E|M)}, \quad \textit{if } |\mathcal{S}| \geq 2. \tag{7.8}$$

The bound for the impersonation attack was first proved by Simmons in 1984 with a long and tedious proof. Several new and much shorter proofs have since then been given. The bound for the substitution attack was proved by Simmons and Brickell, but can also be proved in the same way as for the impersonation attack.

Simmons' bounds give a good feeling of how the authentication protection affects the system. For the impersonation attack, we see that P_I is upper bounded by the mutual information between the message and the key. This means that in order to have a good protection, i.e., P_I small, we must give away a lot of information about the key. On the other hand, in the substitution attack, P_S is lower bounded by the uncertainty about the key when a message has been observed. Thus we cannot waste all the key entropy for protection against the impersonation attack, but some uncertainty about the key must remain for protection against the substitution attack.

Returning to Theorem 7.2, we multiply the two bounds together and get

$$P_I P_S \geq 2^{-I(M;E)-H(E|M)} = 2^{-H(E)}. \tag{7.9}$$

From the inequality $H(E) \leq \log |\mathcal{E}|$ we then obtain the *square root bound*.

Theorem 7.3 (Square root bound)**.** *For any authentication code,*

$$P_D \geq \frac{1}{\sqrt{|\mathcal{E}|}}. \tag{7.10}$$

The square root bound gives a direct relation between the key size and the protection that we can expect to obtain. Thus the following definitions are natural.

An authentication code for which equality holds in the square root bound (7.10) is called a *perfect* A-code. [2] Furthermore, an A-code for which $P_I = P_S$ is called an *equitable* A-code.

Obviously, a perfect A-code must be equitable. If we can construct A-codes for which equality holds in the square root bound we can be satisfied, since in that case no better authentication codes exist, in the sense that P_D cannot be made smaller. This is a main topic, but also equitable A-codes which are not perfect are of interest. The reason for this is the following.

Theorem 7.4. *The square root bound (7.10) can be tight only if*

$$|\mathcal{S}| \leq \sqrt{|\mathcal{E}|} + 1.$$

The square root bound motivates a treatment of non-perfect A-codes, since for perfect A-codes a large source size demands a twice as large key size. This is not very practical. On the other hand, if the source size is very modest, i.e., $|\mathcal{S}| \leq \sqrt{|\mathcal{E}|} + 1$, then we will see in the sequel that perfect A-codes can be constructed for any $P_D = 1/q$, where $q = \sqrt{|\mathcal{E}|}$ is a prime power.

The most important kind of authentication code is when the source message s appear as a part of the channel message m. An A-code for which the map $f : \mathcal{S} \times \mathcal{E} \to \mathcal{M}$ can be written in the form

$$f : \mathcal{S} \times \mathcal{E} \to \mathcal{S} \times \mathcal{Z}, \quad (s, e) \mapsto (s, z), \tag{7.11}$$

where $s \in \mathcal{S}, z \in \mathcal{Z}$, is called a *systematic* (or Cartesian) A-code. The second part z in the message is called the *tag* (or authenticator) and is taken from the tag alphabet $\mathcal{Z}$. We see that systematic A-codes are codes that have no secrecy at all, the source message is transmitted in the clear, and we add some check symbols to it (the tag). In the sequel we study only systematic authentication codes. Systematic A-codes have the following important property.

Theorem 7.5. *For any systematic A-code*

$$P_S \geq P_I. \tag{7.12}$$

[2] The definition of the terminology perfect A-code may be different in other literature.

This means that for systematic A-codes the square root bound is expressed as

$$P_S \geq \frac{1}{\sqrt{|\mathcal{E}|}}.$$

Finally, for systematic A-codes with uniformly distributed keys and $P_I = P_S = 1/|\mathcal{Z}|$, we have the inequality

$$(|\mathcal{Z}| - 1)\,|\mathcal{S}| \leq |\mathcal{E}| - 1. \tag{7.13}$$

This bound shows that large source sizes for equitable A-codes require large key sizes, and gives the motivation for the study of non-equitable A-codes.

We next present some ways of constructing equitable A-codes. Equitable A-codes have the lowest possible probability of deception in the sense that $P_D = P_I = P_S$, but they have the disadvantage of having a source size that is quite modest. It is useful to note that the probability of success in a substitution attack can be written as

$$P_S = \max_{\substack{m,m' \\ m \neq m'}} \frac{|\mathcal{E}(m) \cap \mathcal{E}(m')|}{|\mathcal{E}(m)|}, \tag{7.14}$$

provided that the keys are uniformly distributed.

To have some measure of how good a construction is, we introduce two fundamental definitions. An A-code with fixed parameters $|\mathcal{E}|$, $|\mathcal{M}|$, $|\mathcal{S}|$ and P_I is said to be *weakly optimal* if P_S is the lowest possible. A weakly optimal A-code is said to be *strongly optimal* if, additionally, $|\mathcal{S}|$ has the largest possible value among all the weakly optimal A-codes for fixed parameters $|\mathcal{E}|$, $|\mathcal{M}|$, P_I.

We start by giving the original construction proposed by Gilbert, MacWilliams and Sloane. The construction uses the projective plane.

The projective plane construction: Fix a line L in $\mathbf{PG}(2, \mathbb{F}_q)$. The points on L are regarded as source messages, the points not on L are regarded as keys, and the lines distinct from L are regarded as messages. The mapping from $\mathcal{S}$ to $\mathcal{M}$ means joining the source message s and the key e to the unique line m, which is the resulting message.

We can easily verify the correctness of this construction. The joining of the point e outside L and the point s on L results in a unique line, called m. By running through all pairs (s, e) we find the message space as all lines except L itself. The parameters of the A-code are given by the following theorem.

Theorem 7.6. *The projective plane construction gives parameters*

$$|\mathcal{S}| = q + 1, \quad |\mathcal{M}| = q^2 + q, \quad |\mathcal{E}| = q^2,$$

and the probabilities of success are $P_I = 1/q$ and $P_S = 1/q$.

The A-codes resulting from this construction are strongly optimal.

Another simple construction is the following.

The vector space construction: Let $|\mathcal{S}| = q^m$, $|\mathcal{Z}| = q^m$, and $|\mathcal{E}| = q^{2m}$. Decompose the keys as $e = (e_1, e_2)$, where $s, z, e_1, e_2 \in \mathbb{F}_{q^m}$. For transmission of source message s, generate a message $m = (s, z)$, where

$$z = e_1 + se_2.$$

Theorem 7.7. *The above construction provides* $P_I = P_S = 1/q^m$. *Moreover, it has parameters* $|\mathcal{S}| = q^m$, $|\mathcal{Z}| = q^m$, *and* $|\mathcal{E}| = q^{2m}$.

Authentication codes are closely related to combinatorial designs. This relation has been extensively examined in a number of papers. Brickell, and later Stinson, have established a one-to-one correspondence between A-codes with given parameters and certain combinatorial designs. The designs used are transversal designs, orthogonal arrays, balanced incomplete block designs, perpendicular arrays, etc..

7.3 Constructing useful authentication codes

Our treatment sofar has not considered any results of interest for the case when $|\mathcal{S}|$ is large. This case is of great relevance since many practical problems concern very large source sizes. Examples of such problems are authentication of data files or computer programs. We now turn our attention to the problem of solving the authentication problem for sources that have a length of, say, a million bits. This means $\log |\mathcal{S}| = 10^6$.

So, a fundamental problem in authentication theory is to find A-codes such that $|\mathcal{S}|$ is large while keeping $|\mathcal{E}|$ and P_S as small as possible. In many practical situations one has limitations on $|\mathcal{E}|$ and wants P_S to be bounded by some small value. Also, one usually wants the redundancy ($|\mathcal{Z}|$) to be small, since it occupies a part of the bandwidth.

It has been shown that authentication codes have a close connection to coding theory. It is for example possible to construct authentication codes from error-correcting codes and vice versa. One of the resulting constructions are illustrated next.

We give a construction based on Reed-Solomon codes.

Construction: Let $\mathcal{S} = \{\mathbf{s} = (s_1, \ldots, s_k)\,;\, s_i \in \mathbb{F}_q\}$. Define the source message polynomial to be $s(x) = s_1 x + s_2 x^2 + \cdots + s_k x^k$. Let $\mathcal{E} = \{e = (e_1, e_2)\,;\, e_1, e_2 \in \mathbb{F}_q\}$ and $\mathcal{Z} = \mathbb{F}_q$. For the transmission of source message $\mathbf{s}$, the transmitter sends $\mathbf{s}$ together with the tag

$$z = e_1 + s(e_2).$$

Theorem 7.8. *The construction gives systematic A-codes with parameters*

$$|\mathcal{S}| = q^k, \quad |\mathcal{E}| = q^2, \quad |\mathcal{Z}| = q, \quad P_I = 1/q, \quad P_S = k/q.$$

The construction gives weakly optimal A-codes.