THE PRACTICE OF AUTHENTICATION*

Gustavus J. Simmons
Applied Mathematics Department
Sandia National Laboratories
Albuquerque, New Mexico 87185

Introduction

One of the most pervasive problems in military and in commercial communications-like systems is the need to authenticate digital messages; where authentication is interpreted broadly to mean verification both that a message was originated by the purported transmitter and that it has not been altered subsequently, which includes verifying that it is not a repetition of an earlier legitimate but already accepted message. The terminology "message" is a carryover from the origins of the problem in communications systems, but as used here includes resident computer software, data bank information, access requests and passes or passwords, hand-shaking exchanges between terminals and central facilities or between card readers and teller machines, etc.; i.e., digital information exchange over a suspect channel or interface in general. The need to authenticate information presupposes an opponent(s) -- who may in some circumstances be either the transmitter or receiver -- that desires to have unauthentic messages be accepted by the receiver, or by arbiters, as authentic or else to fraudulently attribute to the transmitter messages that he did not send.

Message authentication is commonly -- and inexplicably -- treated as though it were a single, simple function of the message, much like a parity check, for example. The probable explanation for this is that the result of the authentication operation is two valued; i.e., either the message is accepted as authentic or else it is rejected as inauthentic. What makes the authentication of digital messages such a challenging problem though is that there is no single authentication function that can suffice for all, or even for a large fraction, of the real world authentication needs. For example in some cases the message content cannot (by design) be kept secret from the opponent, while in others he is denied this information. Similarly, in some cases the opponent succeeds if he can cause the receiver to accept any fraudulent message, while in others he succeeds only if he can get a specific message (or one of a small set of messages) accepted as authentic.

* This work performed at Sandia National Laboratories supported by the U. S. Department of Energy under contract no. DE-AC04-76DP00789.

F. Pichler (Ed.): Advances in Cryptology - EUROCRYPT '85, LNCS 219, pp. 261-272, 1986.

In addition, the opponent may either be an outsider with no privileged information or, in a particularly difficult case to deal with, he may be a legitimate receiver misusing insider information provided him so that he can authenticate messages to generate forgeries. Clearly, digital message authentication in circumstances as widely varying as these cannot be expected to be a simple operation. In this paper we shall first derive several theoretical bounds on the quality of authentication that can be achieved for particular authentication channel specifications -- and then exhibit an example system based on the data encryption standard (DES) that seemingly violates these bounds [4]. This apparent contradiction -- arising from the difference between computationally feasible and infeasible problems -- highlights an essential aspect of most practical authentication (or crypto) systems.

Authentication

In message authentication, there are three essential participants: a transmitter who observes an information source*, such as a coin flip, and wishes to communicate these observations to a remotely located receiver over a publicly exposed, noiseless, communications channel; and a receiver who wishes to not only learn what the transmitter has observed but also to assure himself that the communications (messages) that he receives actually came from the transmitter and that no alterations have been made to them subsequent to the transmitter having sent them. The third participant, the opponent, wishes to deceive the receiver into accepting a message that will misinform him as to the state of the source. He can achieve this end in either of two ways: by impersonating the transmitter and getting the receiver to accept a fraudulent message of his own devising when in fact none has been sent by the transmitter, or else by waiting and intercepting a message sent by the transmitter and substituting some other message that is accepted by the receiver as genuine, but which misinforms him of the state of the source. There are many "side" constraints that must be considered in actual applications of authentication:

---

* Ideally we would call the states of the source "messages" as is the practice in communications theory. However, if we did this we would be forced to introduce terminology to designate the collection of sequences that can be sent through the channel, perhaps "authentication code," paralleling "error detecting and correcting code" from communications theory. Unfortunately, the natural contraction "codeword" already has an accepted meaning in communications theory so that we would either have to coin a new word to designate the particular sequence of symbols sent to convey and authenticate a message -- none of which seem very natural -- or else use the cumbersome term "authentication codeword". The term "authenticator", usually used in the sense of an authentication codeword appended to a message, has too restricted a connotation for the general case. We have opted instead to use the term "message' to designate what is actually transmitted and to tolerate the rather artificial device that the information conveyed by a message is the state of a hypothetical source.

the opponent could, for example, be the receiver attempting to falsely attribute messages to the transmitter that he didn't originate and to convince a third party of the authenticity of these forgeries or the transmitter disavowing authentic messages that he did send -- a problem situation commonly referred to as the customer-broker scenario. In the present discussion we only consider one constraint; i.e., whether the receiver is ignorant of the state of the source observed by the legitimate transmitter (authentication with secrecy) or knowledgeable of it (authentication without secrecy).

The source $\mathcal{S}$ is defined by a probability distribution S on its states; s_i. H(S) is the resulting binary entropy of $\mathcal{S}$, i.e., on average this much information must be communicated to the receiver to inform him of the state of the source. If only H(S) bits were communicated, then there would be no possibility of authentication, hence authentication codes -- like error detecting and correcting codes -- depend on the deliberate introduction and use of redundant information in the transmission.

In a common U. S. military authentication protocol both the transmitter and receiver have matched sealed authenticators -- actually a short random sequence of symbols produced and distributed by the National Security Agency. The sealed packets are constructed so as to provide a positive indication (tattle-tale) if they are opened. Each communicant is responsible for the protection of his sealed authenticator and is administratively restrained from opening it until it is needed to authenticate a message. To authenticate a message, the transmitter opens a sealed authenticator, appends the enclosed authentication suffix to the message and then encrypts the resulting extended message. The receiver, upon receiving the cipher would open his matching sealed authenticator and accept the message as genuine if the cipher decrypted to a string of symbols with the proper suffix, and otherwise reject it as inauthentic. If there are r bits of information in the authenticator suffix, an opponent (if he cannot break the "sealing" encryption algorithm) would have only a 2^{-r} probability of "guessing" a cipher which would decrypt into a message ending with the unknown (to him) authentication suffix and hence be accepted as authentic by the receiver. This example illustrates the essential point in all authentication schemes -- namely that for a particular choice of authenticator the receiver will accept only a fraction out of the total number of messages possible.

Continuing the parallel between error detecting and correcting codes and authentication codes, let the particular mapping of states of the source into acceptable (authentic) messages determined by a specific choice (of authenticators) by the transmitter and receiver be called an encoding rule; e_i. $\mathcal{E}$ is then the set of all available (to the transmitter/receiver) encoding rules. Given $\mathcal{E}$, the transmitter and receiver choose encoding rules according to a probability distribution E, which in turn determines a binary entropy of H(E) as to the average equivocation

about the rule being employed. Finally, there is a space $\mathbb{M}$ consisting of all of the messages that the transmitter can use to communicate states of the source to the receiver. We assume that the transmitter can communicate to the receiver any observation he makes of the source; therefore $|\mathbb{M}| > |\mathbb{S}|$ where $|\mathbb{S}|$ is interpreted to be the cardinality of states of $\mathbb{S}$ that have nonzero probability of occurrence. The inequality is strict, since as has already been remarked if $|\mathbb{M}| = |\mathbb{S}|$, all messages would have to be acceptable to the receiver and, hence, no authentication would be possible.

A concise representation of authentication against an opponent impersonating the transmitter is now possible in the form of an $|\mathcal{E}| \times |\mathbb{M}|$ matrix, **A**. The rows of **A** are indexed by encoding rules and the columns by messages. The entry in $a(e_i,m_j)$ is the element of $\mathbb{S}$ encoded by rule e_i into message m_j if such a source mapping exists under e_i and 0 otherwise. As a consequence of the (reasonable) assumption that the transmitter can communicate any state he observes of the source to the receiver, every element of $\mathbb{S}$ appears in each row of **A** at least once and perhaps several times. Earlier comments imply that each row and column contains at least one 0 entry. We now define another $|\mathcal{E}| \times |\mathbb{M}|$ matrix **X**, in which

$$\chi(e_i,m_j) = \begin{cases} 1 & \text{if} \quad a(e_i,m_j) \in \mathbb{S} \\ 0 & \text{otherwise} \end{cases} .$$

For example, for $|\mathbb{S}| = 2$, $|\mathbb{M}| = 4$, the "best" authentication system possible has:

$$\mathbf{A} = \begin{vmatrix} s_1 & s_2 & 0 & 0 \\ s_1 & 0 & s_2 & 0 \\ 0 & s_2 & 0 & s_1 \\ 0 & 0 & s_2 & s_1 \end{vmatrix} \quad \text{and} \quad \mathbf{X} = \begin{vmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{vmatrix} .$$

It is now easy to see the relationship of the impersonation "game" to the matrix **X**. If m_j is an acceptable (authentic) message to the receiver when encoding rule e_i has been agreed to by the transmitter and receiver then $\chi(e_i m_j) = 1$ and the opponent has a probability of success of $p = 1$ if he communicates m_j to the receiver. Conversely, whenever $\chi(e_i,m_j) = 0$ he is certain the message will be rejected. It is certainly plausible -- and in fact rigorously true -- that the opponents probability of success in impersonating the transmitter is the value, v_I, of the zero sum game whose payoff matrix is **X**.

The matrix representation, **Y**, of authentication against an opponent who waits to observe a message sent by the legitimate transmitter and then substitutes some other message is considerably more complex, since the simple strategies available to him in this case are conditional probabilities; i.e., he observes message m_j and must choose, based on this observation, a message m_k to substitute in its stead. The choices available to the transmitter and receiver are the same as in the impersonation case, namely the choice of an encoding rule e_i and perhaps a choice (by the

transmitter) of which message to use to communicate the state of the source, s_ℓ, if e_i maps s_ℓ into two or more messages.

Y is a $|\mathcal{E}| \times |\mathbb{M}|^2$ array whose rows are indexed by e_i and whose columns are indexed by the $|\mathbb{M}|^2$ pairs $m_j|m_k$; meaning m_j is observed and m_k is substituted. The entries in **Y** are:

$$y^i_{jk} = \begin{cases} 1 & \text{if encoding rule } e_i \text{ is used, and when message } m_k \\ & \text{is substituted for } m_j \text{ the receiver is deceived as} \\ & \text{to the state of the source*} \\ 0 & \end{cases} .$$

A strategy for the opponent in this case consists of an $|\mathbb{M}|^2$-tuple (q_{jk}) where $\sum_{k=1}^{|\mathbb{M}|} q_{jk} = 1$, for $1 \le j \le |\mathbb{M}|$, and q_{jk} is the probability of his substituting message m_k for m_j.

Since the opponent has, as a part of his strategy, the option to either impersonate the transmitter before a legitimate message is sent or else to wait and substitute some other message in an attempt to deceive the receiver, the complete authentication "game" is the concatenation of **X** and **Y**, i.e., a rectangular $|\mathcal{E}| \times |\mathbb{M}|(|\mathbb{M}| + 1)$ game subject to the strategy constraints mentioned above plus some additional ones having to do with the transmitter's optimal use of his choices (if any) among available messages to communicate an observed state of the source (splitting). Not surprisingly there exist authentication systems in which the optimal strategy for the opponent is either pure impersonation, pure substitution, optional linear combinations of the two, or most interesting -- essential mixing of both as well as examples in which splitting is essential in the transmitter/receiver's optimal strategies. The point of these remarks is that we have shown in earlier papers that an opponent's overall probability of success under an optimal strategy in deceiving the receiver, P_d, is at least the value of the game whose payoff matrix is the concatenation of **X** and **Y**, and hence that

$$P_d \ge \max(v_I, v_s)$$

where v_I and v_s are the values of the impersonation game on **X** and of the substitution game on **Y**, respectively. We will not pursue this game theory formulation [1] further here, since the purpose of this paper is to review some "channel bounds" on P_d for the authentication channel in terms of the parameters H(S), H(M) and H(E) and then to exhibit a practical system (based on DES) that appears to violate this

---

* If a common state of the source, s_ℓ, is mapped into m_j and m_k under e_i, then even though m_k would be accepted by the receiver as an authentic message when in fact the transmitter sent m_j, he would not be misinformed as to the state of the source and hence $y^i_{jk} = 0$ in this case also.

bound. Of course, the real purpose of the paper is to explain away this apparent contradiction.

The Authentication Channel Bound

Since they have been derived in full detail elsewhere [2], we only exhibit the principal results on the authentication channel bound that are needed to make this paper self-contained. Theorem 1 states the fundamental authentication channel bound.

Theorem 1 (Authentication Channel Bound)

There is a strategy for the opponent such that

$$\log P_d \geq H(MES) - H(E) - H(M) \tag{1}$$

for any transmitter/receiver strategy, E.

Discussion

P_d is the probability that the receiver is deceived as to the state of the source. H(E) and H(M) are the a priori equivocation about the encoding rule being used by the transmitter and receiver and of the message(s) respectively. H(MES) is the joint uncertainty as to the state of the source, the encoding rule and the resulting message. (1) is only one of many equivalent formulations of the channel bound, but is the form most naturally proven by information theoretic arguments.

The expression (1) shows an interesting, and often valuable (to the transmitter/receiver) difference between the essential equivocation faced by the opponent in attempting to deceive the receiver and that faced by the legitimate receiver in attempting to recover the state of the source from the message. The difference in their tasks arises only when the encoding rule encodes at least one state of the source into two or more acceptable messages. In this case H(MES) when expressed in the form

$$H(MES) = H(E|MS) + H(MS)$$

makes clear that to the opponent, meaningful uncertainty about the encoding rule can exist even though he knows both the state of the source and the message chosen to communicate it. H(E|MS) is essentially the uncertainty (to the opponent) introduced by splitting messages. By meaningful, we mean that this uncertainty can be used by the transmitter/receiver to confound him in choosing a substitute message. The receiver on the other hand doesn't care about which message the transmitter chooses,

since he knows the encoding rule and hence will correctly infer the state of the source irrespective of which choice the transmitter made.

Theorem 2 gives eight equivalent expressions to (1) for the case of authentication with secrecy: i.e., the opponent does not know the state of the source observed by the transmitter. This, of course, only matters if the opponent elects to substitute messages rather than to impersonate the transmitter.

Theorem 2

H(MES) - H(E) - H(M) is equivalent to any of the following eight entropy expressions.

	X	Equivalent Form
(2)	ES	$H(M\|ES) + H(S) - H(M)$
(3)	MS	$H(E\|MS) - H(E) + H(MS) - H(M)$
		or
(4)		$H(E\|MS) - H(E) + H(S\|M)$
(5)	ME	$H(E\|M) - H(E)$
		or
(6)		$H(M\|E) - H(M)$
(7)	S	$H(ME\|S) + H(S) - H(E) - H(M)$
(8)	E	$H(MS\|E) - H(M)$
(9)	M	$H(ES\|M) - H(E)$

Proof:

The proof in each case proceeds by splitting the argument in the entropy H(MES) appearing in (1) through conditioning the joint probability on the variable X and then using simple identities to reduce the resulting expressions. The derivation of (2) is typical.

$$H(MES) = H(M|ES) + H(ES)$$

$$= H(M|ES) + H(E|S) + H(S)$$

$$= H(M|ES) + H(E) + H(S)$$

since E and S are independent random variables. Hence

$$H(MES) - H(E) - H(M) = H(M|ES) + H(S) - H(M)$$

as was to be shown, etc.

Eq. (2) says that

$$\log P_d \geq H(S) - H(M) + H(M|ES) \quad .$$

The right-hand term $H(M|ES)$ is the average uncertainty about M given the encoding rule $e_j \in \mathcal{E}$ and state of the source $s_i \in \mathcal{S}$. But there is no uncertainty if there is no splitting, i.e., if there are no choices of messages available to the transmitter to communicate state s_i to the receiver when using encoding rule e_j. In this case, and indeed in general,

(10) $$H(S) < H(M)$$

for authentication to be possible at all. We have noted earlier that the inequality

(11) $$|\mathcal{S}| < |\mathcal{M}|$$

had to also hold for authentication to be possible. In view of the similarity of (10) and (11) a natural question is whether one of the inequalities is stronger than the other, i.e, implies the other. The following small example shows that this is not the case.

Consider two sets $\mathbf{A} = (a_1, a_2)$ and $\mathbf{B} = (b_1, b_2, b_3)$ with probability distributions

$$A = \left(\frac{1}{2}, \frac{1}{2}\right) \quad \text{and} \quad B = \left(\frac{7}{8}, \frac{1}{16}, \frac{1}{16}\right) \quad ,$$

respectively. Then $H(A) = 1$ and

$$H(B) = \frac{25}{8} - \frac{7}{8} \log 7 \approx \frac{2}{3} \quad .$$

Now let $\mathcal{S} = \mathbf{A}$ and $\mathcal{M} = \mathbf{B}$, so that

$$|\mathcal{S}| < |\mathcal{M}|$$

but

$$H(S) > H(M)$$

showing that (11) doesn't imply (10). Conversely, let $\mathcal{S} = \mathbf{B}$ and $\mathcal{M} = \mathbf{A}$, so that

$$H(S) < H(M)$$

then

$$|\mathcal{S}| > |\mathcal{M}|$$

showing that (10) need not imply (11) either. ∎

What is true is that for authentication to be possible both (10) and (11) must be satisfied; i.e., that there must be more messages than states of the source and the average information content in a message must be at least as large as the average uncertainty it resolves (for the receiver) as to the state of the source. Stated in this way, both conditions are certainly reasonably, the only surprising thing is that they are independent.

Using the results of Theorem 2, it is possible to derive some (generally) weaker but enlightening channel bounds. We first note that the total effective equivocation to the opponent playing the substitution game but without knowledge of the source state, i.e., authentication with secrecy, is no greater than $H(E|M)$ and as remarked earlier, the opponent's total effective equivocation if he knows the source state, i.e., authentication without secrecy, is at most $H(E|MS)$.

Theorem 3

For authentication with secrecy

$$\log P_d \geq -\frac{1}{2} H(E) \tag{12}$$

while for authentication without secrecy

$$\log P_d \geq -\frac{1}{2}\{H(E) - H(MS) + H(M)\} = -\frac{1}{2}\{H(E) - H(S|M)\} \tag{13}$$

Proof.

For authentication with secrecy

$$\log P_d \geq \min\{\log v_I, -H(E|M)\} \tag{14}$$

while for authentication without secrecy

$$\log P_d \geq \min\{\log v_I, -H(E|MS)\} \quad . \tag{15}$$

In either (14) or (15) the bounds derived in Theorem 1 and 2 on the value of the impersonation game can be substituted, since the opponent's impersonation strategy is independent of whether he plays substitution with or without secrecy. Replacing

the minimum on the right-hand side of the inequality by the average of the two bracketed terms either weakens the inequality if the terms are not identical or leaves it unaffected if they are. Therefore for authentication with secrecy, replacing v_I with the bound (5) in (14) we get

$$\log P_d \geq \frac{1}{2}\{H(E|M) - H(E) - H(E|M)\} = -\frac{1}{2}H(E)$$

and similarly by replacing v_I with the bounds (3) or (4) in (15) we get

$$\log P_d \geq \frac{1}{2}\{H(E|MS) - H(E) + H(MS) - H(M) - H(E|MS)\}$$

$$= -\frac{1}{2}\{H(E) - H(MS) + H(M)\}$$

or

$$\log P_d = \frac{1}{2}\{H(E) - H(S|M)\}$$

as was to be shown.

Corollary

(16) $$P_d \geq \frac{1}{\sqrt{|\mathcal{E}|}}$$

Proof:

(17) $$H(E) \geq \log|\mathcal{E}|$$

with equality if and only if the transmitter/receiver's optimal strategy E is the uniform probability distribution on $\mathcal{E}$. The conclusion follows by substituting (17) into (12).∎

The expression of the channel bound given in (16) is the one which is apparently violated by the DES based authentication scheme described in the next section.

A "Practical" Authentication Scheme

The source is a "fair" coin flip, i.e., the probability of heads or tails is 1/2. The transmitter/receiver choose to encode (not authenticate) the state "heads" as the sixty-four bit binary string 11...1 and "tails" as 011...1. In other words the redundant information used to authenticate a message is the suffix of 63 1's and only the left-most bit conveys any information about the state of the source. They

then encrypt whichever of these strings is indicated by the coin flip using DES and a secret (known only to them) DES key -- which as is well known consists of fifty-six bits of equivocation to an outsider; the opponent. Each choice of a DES key corresponds in this scheme to a choice of an authentication encoding rule. Consequently, $|\mathcal{E}| = 2^{56}$, and (16) says that

$$P_d \geq \frac{1}{\sqrt{|\mathcal{E}|}} = \frac{1}{2^{28}} \approx 3.7 \times 10^{-9}$$

i.e., the transmitter/receiver cannot, even if they use the 2^{56} encoding rules optimally, limit the opponent's chances of deceiving the receiver to less than roughly four parts in a billion.

Now consider the "practical" chance that the receiver will be deceived. First, there are 2^{64} possible ciphers (messages), only two of which are acceptable for any particular choice of a key (authentication encoding rule). Therefore, if the opponent merely selects a cipher at random and attempts to impersonate the transmitter, his chances of success are 2^{-63} or approximately one chance in 10^{19} not four in 10^{9}! The question is, can he do better. As far as impersonating the transmitter is concerned, the answer is essentially no, even if he has unlimited computing power. For each choice of an encoding rule, there are two (out of 2^{64}) ciphers that will be acceptable as authentic. Assuming that the mapping of $\mathcal{S}$ into $\mathcal{M}$ under DES keys is a random process, this says that the total expected number of acceptable ciphers (over all 2^{56} keys) is $\approx 2^{56.9888}$, i.e., ε close to 2^{57}. Even if the opponent could restrict himself to choosing a cipher from among this collection, his chances of having a fraudulent message be accepted by the receiver would only be 2^{-56} or roughly one chance in 10^{17} which is what we meant when we said that the answer was essentially no. The opponent could not do better, nor worse, (in attempting to impersonate the transmitter) than choose a cipher randomly from among the $\approx 2^{57}$ potentially acceptable ciphers, if the transmitter/receiver chose an encoding rule indexed by the 2^{56} DES keys using a uniform probability distribution as their strategy.

The argument in the preceding paragraph is not misleading, i.e., the opponent's chances of deceiving the receiver through impersonating the transmitter are no better than stated. On the other hand, the channel bound in (16) applies to all authentication schemes -- hence the apparent contradiction must arise in connection with the opponent's substitution strategy. If the opponent waits to observe a legitimate message (cipher), can he put this information to practical use to deceive the receiver. Even if he doesn't know the state of the source, he knows that the cipher is the result of encrypting either 111...1 or 011...1 with one of the 2^{56} DES keys. He also knows that with a probability of essentially one (≈ 0.996), there is only one key that maps the observed message into either of these codes, hence, he is faced with a classical "meet in the middle" cryptanalysis of DES. Clearly if he succeeds in identifying the DES key, i.e., the encoding rule being employed by the

transmitter receiver, he can encrypt the other binary string and be certain of having it accepted and hence be certain of deceiving the receiver. The point, though, is that in order for him to make use of his observation of a message he must be able to cryptanalyze DES. If he can do this, the expected probability of deceiving the receiver is ε close to one -- the small deviation being attributable to the exceedingly small chance that two (or more) DES keys might have encoded source states into the same message (cipher). Thus, we have the paradoxical result that the practical system is some eight or nine orders of magnitude more secure than the theoretical limit simply because it is computationally infeasible for the opponent to carry out in practice what he should be able to do in principle. In this respect practical message authentication [3] is closely akin to practical cryptography where security is equated to the computational infeasibility of inverting from arbitrarily much matching cipher text and plaintext pairs to solve for the unknown key -- even though in principle there is more than enough information available to insure a unique solution.

References

1. G. J. Simmons, "Message Authentication: A Game on Hypergraphs," Proceedings of the 15th Southeastern Conference on Combinatorics, Graph Theory and Computing, Baton Rouge, LA, March 5-8, 1984, pp. 161-192.

2. G. J. Simmons, "Authentication Theory/Coding Theory," Proceedings of Crypto'84, Santa Barbara, CA, August 19-22, 1984, in Advances in Cryptology, Ed. by R. Blakley, Springer-Verlag, Berlin (1985), to appear.

3. G. J. Simmons, "Message Authentication Without Secrecy," in Secure Communications and Asymmetric Cryptosystems, ed. by G. J. Simmons, AAAS Selected Symposia Series, Westview Press, Boulder, CO (1982), pp. 105-139.

4. Data Encryption Standard, FIPS, Pub. 46, National Bureau of Standards, Washington, D.C., January 1977.