2
Cryptography

The Basics

What does it mean to say that communication is secure? In most circumstances, it means that the communication is free from eavesdropping—that the information exchanged is kept private or confidential and does not, in the course of communication, become known to anyone other than the sender and the receiver. The techniques required to achieve this vary substantially, depending on whether the medium of communication is sound, writing, pictures, or some other form of data.

Security can be obtained in a variety of ways. The most common form of secure communication is the private conversation. Although it has become more difficult, in the age of electronics, to be sure of conversational privacy, it is still easier to have privacy in a face-to-face conversation than in any other sort. For a telephone conversation to be private, the speakers must at least have privacy at their respective ends of the line.

The security of conventional handwritten letters is a bit different. It is harder to remain unobserved while reading over the shoulder of someone writing a letter than it is to remain unobserved while listening to someone talk. Unless two people are communicating by passing notes back and forth while sitting in the same room (something people rarely do unless they suspect they are being spied upon), the easiest way to discover what they have written is to intercept the message as it travels from one to the other. In the case of messages written on paper, the primary means of protection is using a trusted means of transport, whether this is a private courier or a state-run mail system. Within this trusted transport medium,

the message is further protected by an envelope, whose function is not so much to prevent entry as to ensure that entry will not go undetected. As we shall see, in the case of electronic messages there is no satisfactory analog to the envelope.

Physical protection is also used to guard electronic messages. A message traveling through copper wires is less vulnerable to interception than one carried by radio. A message traveling through optical fiber is less vulnerable still. Even a message that is sent by radio may be protected by an appropriate choice of frequencies and routes.[1]

There is, however, another possibility. A message may be put at risk of falling into the hands of opponents but may be disguised in such a way that even if unintended parties are able to intercept the message they will not be able to understand it. This is the domain of cryptography.

Less well known than the problem of keeping messages private is the problem of guaranteeing that messages are genuine—of being sure that they really come from the people from whom they appear to have come and that no one has altered them along the way. These properties of communication, called *authenticity* and *integrity*, are arguably more important than privacy. Nonetheless, we will devote more attention to privacy than to authenticity, for several reasons. Although privacy is of limited use in a conversation with someone you do not know,[2] it is generally more difficult to falsify communications than merely to intercept them. Sending a message exposes the sender to discovery in a way that receiving a message does not, because the message invariably exists as evidence of the fact that it was sent. This makes violations of authenticity difficult to achieve under circumstances in which violations of privacy are easy. The foremost reason we focus on privacy, however, is that the right to use cryptography for authentication is not in question; the right to use it for privacy is.

Encrypting a message is often described, by analogy with written messages, as placing it in an envelope, but the analogy is not entirely adequate. A well-encrypted message is far harder to open than one surrounded by paper; however, if the encryption is broken, the break leaves no traces. It is this difference between the functioning of cryptography and the functioning of physical protection mechanisms that gave

rise to the policy issues so prominent in discussions of cryptography in the 1990s.

We shall describe cryptography for the moment only by what it does: a transformation of a message that makes the message incomprehensible to anyone who is not in possession of secret information that is needed to restore the message to its normal *plaintext* or *cleartext* form. The secret information is called the *key*, and its function is very similar to the function of a door key in a lock: it unlocks the message so that the recipient can read it.

The analogy with locks and keys is particularly apt in another respect. The lock and the key are distinct components of a system that controls the use of doors, cabinets, cars, and other things. The lock is a moderately complex mechanical device with numerous moving parts—about two dozen in the case of a normal door lock. The key is a single piece of metal. There are, on the other hand, far fewer types of locks than cuts of keys. Most doors use one of a dozen popular brands of locks, each of which can be keyed to accept one of a million different possible keys. A lock is typically far more expensive than its key, and more expensive to replace, particularly with a lock of a different kind. Perhaps the most important distinction between locks and keys is that locks are not, in principle, secret. Locks are easily recognizable even if they do not display their brand names, and there is no reason to be concerned that people know what type of lock you use on your front door. The cut of the key, on the other hand, is a secret, and any locksmith or burglar who knows it can make a duplicate that will open the door.

In exactly the same way, cryptographic systems are divided into the so-called *general system* (or just *system*) and the *specific key* (or *key*). It has been a principle of cryptography for more than a century (Kerckhoffs 1883) that the general system should not be regarded as secret.[3] The keys, in contrast, must be kept secret, because anyone who is in possession of them will be able to read encrypted messages, just as anyone who is in possession of a door key can open the door.

There is a distinction that is particularly prominent in the older literature between codes and ciphers. *Codes* in this terminology are transformations that replace the words and phrases in a human language with

alphabetic or numeric *code groups*. *Ciphers* are transformations that operate on smaller components of a message, such as bits, bytes, characters, and groups of characters. The distinction is not always entirely clear.[4] Most of the systems we discuss are cipher systems, but codes appear from time to time in historical discussions.

Cryptography in the Small

In using cryptography to achieve secure communication, scale is everything. Two people who meet occasionally and usually communicate by postcard can make use of their infrequent contacts to exchange the secret keys that they will later use to encrypt what they write on their cards. This basic case, in which a small number of correspondents exchange messages of small size, is worth examining in some detail.

Suppose that, as you are about to embark on a journey, a reporter friend asks you for help. Within the next few months there is going to be a demonstration in the city you will be visiting. Your friend has great respect for your powers of investigation and is sure you can learn the time and place of the demonstration. There is one problem. The police are trying to learn just the same information so that they can stop the demonstration. If you call your friend and mention what you have learned, the police will surely overhear since they are tapping all the telephones. What should you do?

Clearly you must encrypt your conversation—encode it in such a way that no one who receives it (except your friend) will be able to understand it. Your friend has told you, however, that the police have all the government's resources available to them. You must, therefore, encrypt your conversation well to keep the police from reading it.

Cryptography on this scale is both theoretically and practically easy. Suppose, for simplicity's sake, that the city you are visiting has a very regular structure with numbered avenues running north-south and numbered streets running east-west. Any address in the city can therefore be given as a pair of numbers, the first representing the avenue and the second representing the street. The time, of course, can be represented by the year, month, day, hour, and minute. (The demonstration is expected to be brief but effective.) Your message will therefore have the form

year month day hour minute avenue street,

where each of these elements is a two-digit number. Perhaps

19 99 12 30 15 25 01 44

means that the demonstration will take place at 3:25 P.M. on December 30, 1999, on the corner of 1st Avenue and 44th Street. Note that every digit in this message is significant. For example, were the demonstration to be only a few days later, four of the digits would change and the value of the year would be 2000 rather than 1999. The digits in some positions, however, are limited in their values. For example, the first digit of the day of the month can only be 0, 1, 2, or 3, and the first digit of the minute can never be more than 5.

In order to encrypt a message of this sort, you and your friend need only agree on a key that will transform the string of 14 digits into some other string of 14 digits. This key must be selected before your departure, and you must carry it with you and keep it secret.

In order to carry out the transformation you will add the digits of the message one at a time to the digits of the key, without carrying.

Suppose that the key is

64 25 83 09 76 23 55 72

and you add the message

19 99 12 30 15 25 01 44.

You will get

73 14 95 39 81 48 56 16

as the cryptogram. Note that in some cases the addition produced a number greater than 10. In the third place, for example, the sum is 11. In such cases the 1 in the tens place is simply thrown away, rather than "carried" into the next place as in ordinary arithmetic.

Once you have learned the time and place of the demonstration, you convey your message by calling and reading the encrypted version over the phone.[5]

To decrypt the received string of numbers, your reporter friend will take the cryptogram

73 14 95 39 81 48 56 16

and subtract (without "borrowing") exactly the same numbers you added:

64 25 83 09 76 23 55 72.

This will yield the original message:

19 99 12 30 15 25 01 44.

Assume that the police have intercepted your phone call. No matter what computers or codebreaking skills they may possess, they have no hope of recovering the underlying message; there simply is not enough information in what they have received. The fact that the key was chosen entirely at random means that for any possible message there is a key that would produce any observed cryptogram. For example, if the time and place of the demonstration had instead been January 11, 2000 at 5th Avenue and 23rd Street

20 00 01 11 10 45 05 23

and the key had been:

53 14 94 28 71 03 51 93

the cryptogram would have come out exactly the same.

A cryptosystem of this kind is called a *one-time* system because it is perfectly secure if used only once. If it is used even twice, the results are likely to be disastrous.

Suppose that, in your travels, you learned about not one demonstration but two. Since you have brought only one key along on your trip, you use it for both messages. Unfortunately for the second demonstration, the police cryptanalysts figure out what is happening. When the first demonstration occurs and, despite the secrecy of its planning, gets mysteriously good coverage in the foreign press, they consider the demonstration in light of your two telephone calls. By combining your first message with the date of the first demonstration, they extract what they presume to be a key. When they decrypt your second message with the same key, they get another place and time[6]; with that information, they can prevent the second demonstration before it begins.

If you follow sound procedures, this sort of error will never occur. You might perhaps carry only one copy of the key, written on the paper of a cigarette. Once the message has been sent, you can light the cigarette in the lee of the telephone booth and stroll off down the street, feeling like a real spy.

Cryptography is, of course, not limited to the occasional exchange of short messages between friends. It may require millions of bits of information to be encrypted in order to protect a first-run movie or a semiconductor mask file.[7] It may require the exchange of messages among hundreds or thousands of people to protect the communications of a large corporation. Often it requires both. As cryptography grows in each of these directions, it rapidly becomes more complex.

One-Time Systems on a Larger Scale

The scenario that has just been described is entirely practical. In fact, the same procedure can be used and has been used on a much larger scale. It was a mainstay of Soviet diplomatic communications from the late 1920s until at least the early 1950s. Since their messages did not always have the convenient numerical character of those presented in our example, the Soviets first had to convert them into numerical form. This was done with *one-part codes*[8] similar to the following:

abovementioned	0000
academician	0001
acknowledge receipt	0002
arrange meeting	0003
avoid contact	0004
...	...

These four-digit code groups were then added to four-digit key groups in the same digit-by-digit fashion employed in our example above.

Using one-time systems in a large network creates a number of serious problems. First, although it is trivial to produce a few dozen or a few hundred random digits by throwing 20-sided dice (figure 2.1) from a fantasy games shop, it is quite another thing to manufacture millions upon

Figure 2.1
Twenty-sided dice. (Photograph by Eric Neilsen.)

millions, type them up onto sheets, and produce precisely two copies of each sheet.

Before we go further, a word about the terminology of modern cryptography is in order. In many papers, the participants in encrypted communication are personified, whether they are people, pieces of equipment, or computer processes. If there are two parties involved, they are called Alice and Bob. If more are required, they are drawn from a cast of supporting characters that includes the couple's friends Carol and Ted, along with Eve, the eavesdropper.

In a network with more than two correspondents, there is difficulty in coordinating the keys used. If three people share a body of one-time key and Alice uses some in sending a message to Bob, she must inform Carol of what she has done. If Carol does not know this, she will at some time use the same key to send a message of her own and thereby create an insecurity. The feasibility of such coordination among three people is clear; for 1000, it is not.

The way the Soviets dealt with this problem was by having all communications go through Moscow. Every embassy could communicate

securely with Moscow using keys that it shared only with Moscow. If the Soviet embassy in the United States needed to communicate securely with the Soviet embassy in Mexico, it was required to send its message to Moscow and have it relayed to Mexico City. Such an arrangement makes a network less flexible and requires twice as much keying material to be expended in sending each message.

Despite the centralized approach, the Soviets got into trouble. For reasons that are still unclear, a serious mistake was made in the early months of 1942. Rather than making exactly two copies of the key sheets, they made four. These excess keys then entered the inventory and remained in use for several years. Western intelligence noted and exploited the multiple use of the keys, with disastrous results for Soviet security. Under the code name Venona, cryptanalytic study of the reused "one-time" keys went on for decades. The system was used for the most sensitive Soviet information, and the Americans and the British studied it in hopes of identifying Soviet "moles" thought to be operating at the highest levels of their intelligence establishments.[9]

One-time systems are not the only form of highly secure cryptography, and they are by no means the dominant form today. In order to avoid having to ship the titanic amounts of keying material that are required in one-time systems, most enciphering today is done by *cipher machines*: mechanical or electrical or computer devices that encode messages. One-time systems have the advantage of simplicity but the disadvantage of failing completely if used to encrypt an amount of text exceeding the size of the key. The functioning of cipher machines is more complex than that of one-time systems. This complexity is the price of a system that can protect quantities of traffic far greater than the size of the key.

A Brief History of Cryptographic Systems

Despite the vast progress of cryptography during the twentieth century, there is a remarkable continuity with systems that have been known since the Renaissance.

Cryptography is always a matter of substituting one thing for another. The earliest cryptographic systems substituted one letter of the alphabet for another in an unchanging fashion, a technique called a *simple* or

```
b o o k k e e p e r

g a a o o b b t b w
z e e w w i i a i k
o z z e e j j s j y
n s s i i r r o r q
```

Figure 2.2
The characteristic letter pattern of the word 'bookkeeper'.

monoliteral substitution. Simple substitutions are easy to perform, even when the computational resources are limited to pencil and paper. There are also plenty of them: some 2^{90} for a 26-character alphabet. In other words, a simple substitution cipher has a 90-bit key—far larger than anyone could have needed before the computer age. Despite these virtues, simple substitution ciphers are quite easy to break. This is because they leave many characteristics of the message, such as letter frequency and letter patterns, unchanged.

The best-known approach to solving simple substitution ciphers is to compute the frequencies of the various letters. In English, the letters of the alphabet have widely varying rates of occurrence. The letters E and T, for example, occur quite frequently, accounting for 13% and 9% of typical text, whereas J and Z account for only 2% and 1% of such text. These characteristics frequencies permit a cryptanalyst to recognize the identities of the letters despite the substitution. Analysts also make use of the preservation of letter patterns. Figure 2.2 shows that the exceptional structure of the word 'bookkeeper' remains visible when it is encrypted under a variety of cipher alphabets. Notice that in each case the resulting cryptogram shows the letter pattern 1223344536—that is the cryptogram contains three consecutive pairs of repeated letters in the middle. Admittedly, this is a word chosen for its exceptional pattern of repeated letters (it is the only English word with three pairs of repeated letters in a row); however, it is only an extreme case of a very common phenomenon that occurs in such words as 'pepper', 'papa', and 'noon'. Repetition patterns allow a skilled cryptanalyst to read words of this sort directly from the ciphertext.

The cure for the shortcomings of monoalphabetic substitution—

discovered some 500 years ago and still in use today—is to change the cipher alphabet from one letter to the next. *Polyalphabetic encryption*, as this is called, is the creation of three Renaissance scholars, Alberti, Belaso, and Trithemius (see Kahn 1967), but is commonly known by the name of another, Blaise de Vigenère—an error too deeply embedded in cryptographic terminology to admit of historical correction at this date.

The simplest form of polyalphabetic cipher employs a sequence of alphabets. The first alphabet is used to encrypt the first letter of the message, the second alphabet to encrypt the second letter, and so on. Once the supply of alphabets has been exhausted, the encipherer starts over again with the first. The cipher alphabets may either be unrelated, as in figure 2.3,[10] or may be generated by simple transformations from a single alphabet. The more distinct alphabets are used, the more secure the system, but the more it suffers from the problems of a one-time system—the amount of keying material becomes excessive.

The general form of the Vigenère system employs a number of independent cipher alphabets, as in figure 2.3, and employs some subset of them sequentially in a pattern that may or may not repeat and may or may not use all of the alphabets. Note that each alphabet is labeled at the left with a letter of the alphabet. This allows a *key word* or *key phrase* to represent a sequence of alphabets, as in figure 2.3.

Polyalphabetic ciphers call on three basic processes to achieve security: they employ a set of unrelated cipher alphabets, they derive a number of secondary alphabets from each of these primary alphabets, and they vary the use of the secondary alphabets in a more or less complex pattern. These three processes can be traded off against each other. The more complex the pattern in which the alphabets are used, the smaller the number of distinct alphabets that are needed. Of the innumerable variations of polyalphabetic ciphers that are possible, we will examine a small number of examples illustrative of the development of cryptography from the Renaissance to the twentieth century.

The simplest form of Vigenère cipher uses *direct standard alphabets*—that is to say, the ordinary alphabet (standard) in its usual order (direct).[11]

The difficulty of solving a Vigenère system is entirely dependent upon the relationship between the length of the key and the length of the

Vigenère table

	A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
A	Q F A L H I M Z E T Y N B O U D X P C S K G R J V W
B	Y L K O R U C J X P A S V H B D Q G M I T E Z J W N
C	J O A M S I T Y R D N H X E W P F V Z B L G K Q U C
⋮	. . .

Plain:	d	o	d	e	c	a	h	e	d	r	o	n
Key:	*B*	*A*	*F*	*F*	*L*	*E*	*D*	*B*	*A*	*F*	*F*	*L*
Cipher:	O	U	A	S	M	O	P	R	L	U	Q	Y

"d" is carried to "O" under the "B" alphabet;
"o" is carried to "U" under the "A" alphabet;
"d" is carried to "A" under the "F" alphabet; etc.

Figure 2.3
Action of a Vigenère system.

message to be encrypted. After a certain point (called the *period*) in the encryption, a new sequence of message characters will be encrypted with the same sequence of key characters. The number of times this occurs is referred to as the *depth* of the message with respect to the key. The greater the depth, the easier the message is to break.

Despite the fact that Vigenère systems with short periods are not secure, they are often used. One example is the CAVE cipher still used to protect digital cellular telephone calls in the United States (Electronic Industries Association 1992).[12]

A little examination will show that the short "time and place" message with which we began was encrypted in a Vigenère system using numbers instead of letters. A central theme in cryptography has been to produce systems with long periods without having to transport the large amounts of keying material required by the one-time systems. One way of solving the problem of short periods is to encrypt the message more than once—a *multiple Vigenère system* (figure 2.4).

Although encrypting two, three, or more times makes for a vast improvement in the security of messages, it was not actually feasible when enciphering was done by hand. In practice, errors by the code clerks—

```
Plain:      HENRY IS HUNGRY ...
Key-1:      PAPAD UM PAPADU ...
Cipher-1:   WECRB CE WUCGUS ...
Key-2:      HIPSH IP SHIPSH ...
Cipher-2:   DMRJI KT OBKVMZ ...
```

Figure 2.4
Double Vigenère system.

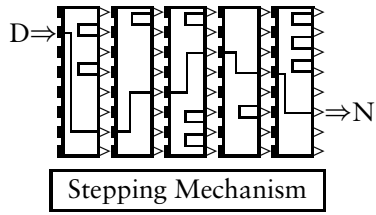

Figure 2.5
Rotor machine.

possibly in both the enciphering and the deciphering—made many messages unreadable. The wide use of such complex encryption techniques only appeared in the twentieth century, with the development of electromechanical enciphering equipment. The first of these was the *rotor machine*, a device that was to dominate cryptography for half a century.

The central component of a rotor machine is the *rotor*, a disk about the size of a hockey puck that serves to implement a cipher alphabet. On each face of the disk there are a number of electrical contacts corresponding to the letters of the alphabet. Each contact on the front face is wired to exactly one contact on the rear face. As an electrical signal passes through the rotor, the signal is carried to a new alphabetic position, just as a letter looked up in a cipher alphabet changes to another letter.

Rotor machines have had from three to more than ten rotors. Every rotor through which a "letter" passes represents an additional layer of encryption. Thus, the simplest rotor machines correspond to triple Vigenère systems, and the more elaborate ones may be several times as complex.

With the appearance of digital electronics after World War II, the dominance of rotor machines gave way to that of *shift registers*. A shift

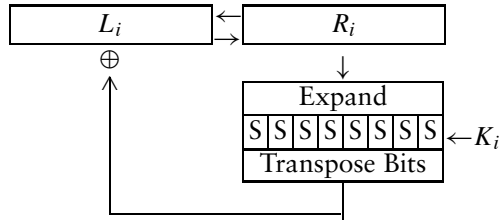

Figure 2.6
DES as a shift register.

register is an electronic device made up of a number of *cells* or *stages*, each of which holds a single 0 or 1 of information. As the shift register operates, the data shift one or more places along the register at each tick of the clock. In addition to moving left or right, some of the bits are modified by being combined with other bits. In the more modern *nonlinear* shift registers, some of the bits are looked up in tables[13] and then used to change other bits of the register, under control of a key. This process is repeated over and over until every bit has changed in a way that is a complex function of every other bit and of every bit of the key. In other words, if a single bit of either the input or the key is modified, approximately half of the output bits will change.

Typical of modern non-linear shift-register systems is the US Data Encryption Standard (USDoC 1977). DES, as it is generally known, is a *block cipher* or *electronic code book*. It takes 64 bits (8 bytes) of information as input, and, under the control of a slightly smaller (56-bit) key, produces 64 bits of output. DES is rarely operated in the basic mode in which it is described in the standard. It is intended as a primitive for building more complex modes of operation suitable for use in encrypting common data formats (USDoC 1980).

DES is a shift-register system in a modern style. In each of its 16 rounds it performs a complex operation on half of the 64 bits in its register and uses the result to alter the other half of its bits. Each round is controlled by a distinct 48-bit subset of the key.

DES expands the right-hand side of its register from 32 to 48 bits and exclusive-ors these with 48 bits selected from the key. It then contracts the 48 bits back to 32 by making eight substitutions, using six-bit-to-

four-bit tables. These tables, called *S-boxes* (for *selection boxes*), are the heart of the algorithm; along with the key size, they have been sources of controversy (Bayh 1978). Finally the bits are rearranged (*transposed*) before the result is exclusive-ored with the left side of the register.

A proposed replacement for DES was Skipjack, an NSA-designed algorithm with a 64-bit blocklength and an 80-bit key.[14] Skipjack operates on its text 16 bits at a time and employs a number to techniques not present in DES. It has two types of rounds, doing eight of one then eight of the other. It has only one S-box but one equal in size to the sum of DES's eight, which it varies from occurrence to occurrence by adding in a counter. It has a simple key schedule compared with DES's complex one. Overall, Skipjack is a cleaner and more attractive algorithm than DES that might have been successful as a DES replacement had it not been introduced as part of a plan for *key escrow*. Key escrow is a mechanism for guaranteeing that some third party—in Skipjack's case the US government—is always able to read the encrypted traffic. To this end, the Skipjack algorithm was kept secret for six years to prevent unauthorized *unescrowed* implementations. When the escrow program had clearly failed, Skipjack was declassified to allow its use in software for email protection on military networks. It was too late, however, for the algorithm to gain acceptance as a replacement standard. A contest to select such a replacement was already underway; the system it produced will be discussed shortly.

Just as rotor machines look at first to be quite different from Vigenère encipherment, shift registers look superficially quite different from rotor machines. There is, however, a deep similarity. All these systems combine a process of looking things up in tables with one of "adding" them together using some sort of arithmetic operation. In shift registers, the message and the key are more thoroughly mixed together, rather as though the positions of rotors in a rotor machine were affected by the letters passing through them.

Strengths of Cryptosystems

Before we go further into cryptosystems and their use, a word about the strengths of cryptosystems is in order. A cryptosystem is considered secure when an opponent cannot break it under reasonable circumstances, in a reasonable amount of time, at a reasonable cost. The term "reasonable" is perforce vague. Despite being the most important problem in cryptography, the evaluation of cryptographic systems is the least understood. An adequate mathematical theory of cryptography has eluded cryptographers for centuries.[15]

The issue of what constitutes *reasonable* conditions for an attack on a cryptosystem is better understood than others. An opponent must, of course, be in possession of ciphertext to have any hope of discovering the underlying plaintext. If this is all that is available, we say that the cryptanalyst mounts a *ciphertext-only attack*.[16] Typically, however, the opponent knows some information about the plaintext before starting to work on the problem. A message from Alice to Bob, for example, is likely to begin "Dear Bob" and to be signed "Love, Alice." Although knowledge of such *probable words* or *cribs* is difficult to prevent, the situation may go much farther than this. Many messages, such as product announcements and press releases, are secret until a certain date and then become public. Under these circumstances an opponent has the corresponding plain text and cipher text of one message and can make use of this in attacking other messages sent using the same key. If this is the case, we say that the opponent is in a position to mount a *known-plaintext* attack.

At the time of World War II, the belligerents, not trusting their cryptosystems to resist known-plaintext attacks, imposed such *signaling rules* as "No message transmitted in cipher may ever be sent in clear." In fact a message sent in cipher could never be declassified without being *paraphrased* to reduce its cryptanalytic utility. This problem appears to have been solved for US government cryptosystems by the 1960s, permitting formerly encrypted messages to be declassified on the basis of content alone.

The sort of signaling rules formerly used by the military are entirely

infeasible in a commercial environment. Fortunately, trust in modern electronic cryptosystems is sufficient that the availability of plaintext to an opponent makes no difference. In fact, we presume that the opponent can send an arbitrary number of text messages and will receive our cooperation in enciphering them or deciphering them before beginning to attack the actual message of interest. This is called the *chosen-plaintext* assumption.[17]

Workfactors

Time is the essential element in measuring computation. The question is "How much will it cost me to get the answer when I need it?" It is a rule of thumb that computing power doubles every 18 months[18]; thus a personal computer purchased for $1000 today will have twice the computing power of one purchased less than 2 years ago.[19] These improvements in speed have profound implications for cryptographic systems.

The number of operations required to break a cryptographic system is called its *workfactor*. The form or complexity of the operations is not precisely stated. They might be encryptions, as they are when the analytic process is one of searching through the keys, or they might be something entirely different. In a spirit of oversimplification, we will assume that operations are always entirely parallelizable. If two processors can do a million operations in 5 seconds, then ten processors can do the same number of operations in 1 second. For our purposes, this assumption is conservative in the sense that if it is false the problem merely becomes somewhat harder.

If a system has a workfactor of 2^{30}, it can be broken by a billion operations. These may not be elementary computer instructions; they may be complex operations requiring hundreds of instructions each. Even so, typical desktop computers today can do a billion instructions a second. If such a cryptosystem requires several hundred instructions per encryption, it can be searched in minutes. In short, breaking a system with a workfactor of 2^{30} is trivial.

A workfactor of 2^{60} means that a million processors, each doing a million operations a second, can solve the problem in a million seconds (between 11 and 12 days). It is clear that a system with a workfactor of

2^{60} can be broken today if the analytic operations are such that processors capable of executing them are worth building or already available. If the operations are encryptions, the processors might be built from available encryption processors.

On this path, systems with workfactors of 2^{90} are the first that seem beyond reach for the foreseeable future. A billion processors in parallel can certainly be imagined. A billion operations a second, even operations as complex as DES encryptions, had already been achieved around 1990 (Eberle 1992). A billion seconds, however, is 30 years—long enough to count as secure for most applications.[20]

Workfactors of 2^{120} seem beyond reach for the indefinite future. A trillionth of a second is less than one gate delay in the fastest experimental technologies; a trillion processors operating in parallel is beyond reach; a trillion seconds is 30,000 years.

The only technological development on the horizon that would be capable of bringing such computations within reach is *quantum computing*. Quantum computing makes use of *superposition* of physical states to calculate all of a set of possibilities simultaneously. To date, its application to any real problem remains a fiction. Quantum computing has the potential to destroy all of the public-key cryptosystems currently in use and to cut the effective key lengths of conventional cryptographic systems in half. Quantum computing, however, is unlikely to be an immediate threat. Breaking the most secure key management systems in use today would require factoring 2000-bit numbers; quantum computing made news when it factored the number 15 (Vandersypen et al. 2001).

Estimating the cost of searching through keys and validating the estimates by actually doing it was a sport in the cryptographic community for some time. In the fall of 1995 a group of cryptographers met and prepared an estimate of search costs, concluding that 40-bit keys (the largest that could be readily exported at the time) could easily be searched and that keys at least 70 to 90 bits long were needed to provide security for commercial applications (Blaze et al. 1996). The previous August, students at the École Polytechnique in Paris had searched out a 40-bit key. The following January, students at MIT repeated the feat using an $83,000 graphics computer. This amounted to a cost of $584 per key. At its annual conference in January 1997, RSA Data Security offered prizes

for searching keyspaces of various sizes. The 40-bit prize was claimed before the conference ended and the 48-bit prize was claimed a week later. The 56-bit DES challenge lasted for only 5 months.

The US Data Encryption Standard used a 56-bit key and thus falls within the range we have described as clearly possible. The standard has been used extensively throughout the commercial world—particularly by banks, which commonly engage in billion-dollar electronic funds transfers. In such applications, the inadequacy of any algorithm with a 56-bit key is apparent. Because the National Institute of Standards and Technology (NIST) was slow to issue a replacement standard, *triple-DES*—a block cipher employing DES three times in a row with three different keys[21]—arose as a de facto standard and was formally adopted first by the Banking Security Standards Committee (ANSI X9F) of the American National Standards Institute (ANSI 1998) and the National Institute of Standards and Technology (FIPS 46-3).

Eventually, DES was replaced by a new cipher using much longer keys. This system, which is called the Advanced Encryption Standard or AES and will be discussed shortly, has largely ended the game of searching keyspaces.

Lifetimes of Cryptosystems

In designing a cryptographic system, there are two important lifetime issues to consider: how long the system will be in use and how long the messages it encrypts will remain secret.

Cryptographic systems and cryptographic equipment often have very long lifetimes. The Sigaba system, introduced before World War II, was in use until the early 1960s. The KL-7, a later rotor machine, served from the 1950s to the 1980s. DES was a US standard for some 25 years. It is still in widespread use, and it may be for decades.[22] Other systems that are neither formal standards nor under the tight control of organizations such as the American military probably have longer lifetimes still.[23]

Secrets can also have very long lifetimes. The Venona messages were studied for nearly 40 years in hopes that they would reveal the identities of spies who had been young men in the 1930s and who might have been the senior intelligence officers of the 1970s. The principles of the Sigaba system were discovered in the mid 1930s and were not made

public until 1996. Much of the "H-bomb secret" has been kept since its discovery in 1950, and the trade secrets of many industrial processes are much older. In the United States, census data, income tax returns, medical records, and other personal information are supposed to be kept secret for a lifetime.

If we had set out to develop a piece of cryptographic equipment in the late 1990s, we might have expected it to be in widespread use today. We might also reasonably plan for the system to stay in use for 25 years or more. No individual piece of equipment is likely to last that long; however, if the product is successful, the standards it implements will. If the equipment is intended for the protection of a broad range of business communications, some of the messages it encrypts may be intended to remain secret for decades. The cryptosystem embodied in our equipment might thus encrypt its last message in 2030 or later, and that message might be expected to remain secret for 25 years more. The system must therefore withstand attack by a cryptanalyst in the late twenty-first century, whose mathematical and computing resources we have no way of predicting. The prospect is daunting.

The Advanced Encryption Standard

The daunting prospect was taken on by a process begun in 1997 and concluded in 2001, when the Data Encryption Standard was replaced by the Advanced Encryption Standard. The new system is an improvement over the old in every respect. It doubled the length of the block from 64 to 128 bits, increased the size of the key to between 128 and 256 bits, and substituted a mathematically based design for one dominated by engineering concerns.

No mathematical theory behind the tables at the center of DES was included in the standard or in any accompanying material. More important, no significant cryptanalytic techniques that might be applied to DES were publicly known when it appeared. This reduced any attempt to prove that DES was secure to vague generalities. If we could develop an algorithm based on a mathematical theory of the cryptanalysis of block ciphers, we could have proofs that the algorithm would resist certain types of attacks. If the attacks were sufficiently general in scope, resistance to the attacks might reasonably be described as security.

The starting point for modern cryptography was put forth by Claude Shannon, the founder of information theory, in 1949. He proposed combining *confusion* (intrinsically complex mathematical operations that perforce operate on small quantities of data) with *diffusion* (operations that spread the effects of confusion across larger data elements). In DES, the confusion was provided by a set of lookup tables and the diffusion by permutations of bits.

Beginning in the late 1980s, cryptographers began applying algebraic techniques to improve both components. These theories were based on the theory of finite fields. Slightly later two fundamental cryptanalytic techniques were developed. *Differential cryptanalysis* analyzes the pattern of changes in output resulting from changes in input; *linear cryptanalysis* makes use of a deep mathematical fact—that no transformation can be purely non-linear—to derive expressions approximating the key bits. This made it possible to give proofs that systems built using algebraic structures would resist differential and linear cryptanalysis.

Like much of cryptographic work, the research took two steps forward, and then a step back. It produced the algorithm SHARK, which had good *diffusion* (spreading the attacker's attention over large numbers of bits), but a plaintext/ciphertext attack on a simplified version of SHARK showed other problems. This led to development of the algorithm SQUARE, which in turn succumbed to attacks on its byte-oriented structure. Further improvements let to the algorithm Rijndael, which became the Advanced Encryption Standard.[24] Rijndael[25] operates approximately as follows. The input is a 128-bit block organized as a 4×4 matrix of 8-bit bytes. In each round of the Rijndael algorithm, there are four steps:

- Add in the key.
- Look the bytes up in a table.
- Rotate the rows.
- Apply a linear transformation to each column. (Landau 2004, p. 108)

Rijndael's confusion step is the table lookup; its diffusion steps are the row and column operations. The combination, done ten to fourteen times

(depending on the size of the key), provides the algorithm's security; the mathematical formulation of the algorithm provides a base from which to analyze that security.

Key Management

Key management—the production, shipment, inventorying, auditing, and destruction of cryptographic keys—is an indispensable component of secure communication systems. Cipher machines make a spectacular reduction in the amount of keying material that users must ship around. This diminishes the problem of key distribution; however, it does not eliminate it, since the difficulty of distributing keys is typically more a function of how many people are involved than of whether each one has to get a large codebook or a short message.

The production of keys is the most sensitive operation in cryptography. Cryptographic keys must be kept secret and must be impossible for an opponent to predict. If they do not achieve these objectives, the results will be disastrous, regardless of how cleverly designed the cryptographic systems in which they are used. The failure of Soviet key production that led to the breaking of the Venona intercepts is one example of this. A far more recent example is the penetration of the Secure Socket Layer (SSL) protocol, which is used to secure Internet transactions by encrypting such sensitive information as credit card numbers. In the summer of 1995, a group of graduate students at the University of California at Berkeley discovered that the SSL protocol generated session keys by consulting the clock on the client machine (typically a personal computer). The time on the client machine's clock could be inferred quite accurately from other aspects of the protocol, and thus the key could be discovered easily.

In conventional cryptographic practice, reliability in key production is achieved by centralization. If all communications go to and from a central site, as in the Soviet diplomatic network, it is natural to manufacture keys at the center and ship them to the far-flung sites. The same procedure is typically followed even in cases where messages can flow directly from one field organization to another.

In US parlance, the organization that manufactures keys is called the *central facility*. The production process goes to great lengths to guarantee

that the keys produced are completely random and are kept completely secret. From the central facility, keys are shipped to installations around the world through a special administrative structure called the COMSEC *materials control system*, which uses such elaborate security procedures as *two-person control*.[26] Keying material is stored by COMSEC *custodians* and constantly tracked by an elaborate system of receipts.

At one time, keys took the form of codebooks. Originally, these were produced by writing the words and phrases of the plain text on a set of index cards and shuffling the cards by hand. In the 1930s, handwritten or typed index cards were replaced by punched cards, and the shuffling was accomplished by using card sorters to sort cards at random. Later cryptographic systems used wired rotors, key lists, plugboards, and paper tape. The most recent have keys packaged in entirely electronic form.

Because keys are secret, they must have names to provide the users with a way of "talking" about which ones to use. A message sometimes contains the name of the key that is to be used in decrypting it. At other times the key is a function of the message's origin or subject matter. Thus an embassy might have one system for diplomatic messages, another for messages dealing with trade, and yet a third for the messages of the military attaché. A ship at sea might have one set of keys for communicating with shore-based facilities and another for communicating with each of the fleets with which it was likely to come in contact.

The distribution of keys follows the structure of the organization that employs them. Keys, however, usually change more often than organizational structures, and it would be confusing to ask a code clerk to remember an entirely new key name when the key was used in exactly the same way. The solution is to keep the name constant but to label each new key as to its *edition* (a term that dates from the era of codebooks but is still used today in naming purely electronic keys).[27]

The management of keys is a constant tug of war between the need for flexibility in communication and the need to maintain security by limiting keys. Typical of military communications is the US Navy's Fleet Broadcast System, which transmits constantly to American ships all over the world. The keys used by the Fleet Broadcast System are changed every day, but on any given day they are the same for every ship. This arrangement favors flexibility over security and has had its costs. In

the late 1980s, a Navy chief petty officer named Jerry Whitworth sold keys from the Alameda Naval Air Station to Soviet intelligence officers. Because of the widespread distribution of those keys, the USSR was potentially able to decipher many US Navy communications. At the other extreme, keys may be issued for a communication network with only two members. The result is more secure but less flexible. If keys are distributed physically, the endpoints of the circuit must anticipate the need to communicate far enough in advance to order their keys—a process that may take days or weeks. In many circumstances (for example a secure phone system), this is an unacceptable burden and some less cumbersome solution must be found.

Dynamic Key Distribution

Suppose, in the traditional terminology of the modern cryptographic literature, that Bob wants to communicate with Alice. Bob may not know Alice; he may simply know that she is a lawyer, an investigator, or a doctor whose expertise and confidential collaboration he requires. If the proper arrangements have been made in advance, Bob and Alice can be "introduced" to each other in real time.

The mechanism that makes the introduction is called a *key management facility* (KMF).[28] It is a network resource, similar in function to a directory server, that shares a key with every member of the network.[29] In the description that follows, we will assume that Bob is using a cryptographically secure telephone to call Alice, and to simplify the description we will blur the distinction between the people and the instruments.

If Bob and Alice belong to the same community and rely on the same KMF, the process goes as follows:

- Bob calls the KMF and informs it that he wants to communicate with Alice. This call is encrypted in a key that Bob shares with the KMF.
- If the KMF finds that it "knows" Alice (i.e., that it shares a key with her), it manufactures a new key for the exclusive use of Bob and Alice and sends it to each party. In the particular case of telephones, which can have only one call in progress at a time, this is done by sending Bob a two-part message. Both halves of the message

contain the same key, but the first half of the message is encrypted so that Bob can read it and the second half so that Alice can read it (Needham 1978; Rosenblum 1980).

- With the key now in hand, Bob calls Alice. The portion of the key that was encrypted so that only Alice can read it functions as a letter of introduction; Bob's phone sends it to Alice's phone at the beginning of the call.
- Bob's telephone and Alice's telephone now *hold* a key in common and can use it to make their secure phone call.

The *long-term* keys that Bob and Alice share with the KMF may reasonably be called *subscriber keys*; the ones they use for particular messages or conversations are called *session keys* or *traffic keys*. Because of the extra phone call required for Bob and Alice to acquire a common key, systems of this sort typically *cache* keys for some period of time. This allows the users to make repeated calls without the overhead of the KMF call.

This approach to keying secure phone calls is far from ideal. For one thing, if communication between any two members of a community requires the services of the KMF, it will be a busy facility indeed.[30] A far more serious problem, peculiar to security, is that Bob and Alice must place too much trust in the KMF.

Strengths and Weaknesses of Cryptography

Cryptography is the only technique capable of providing security to messages transmitted over channels entirely out of the control of either the sender or the receiver. To put this another way: the cost and applicability of cryptography depend very little on either the length or the "shape" of the path the encrypted message must follow.

Cryptography is not always the most appropriate security technique. In designing a system for securing the communications within a building, a campus, or a military base, it is appropriate to protect the signals physically by running them through shielded conduits or optical fibers. On the other hand, if the trunks of a network span great distances, cryptography is usually the most economical security mechanism and in

some cases the only possible one. Physical protection techniques that are perfectly suitable for a network spanning distances of a few hundred yards will be prohibitively expensive if used to connect cities as far apart as Paris, Los Angeles, Hong Kong, and Sydney. If cryptography is used to provide security, however, the cost will be independent of the distance. Mobile and satellite communications represent an extreme case in which cryptography appears to be the only possible means of protection.

From a managerial viewpoint, security obtained through the use of physical means of protection must be bought at the cost of constant auditing of the signal path. If land lines are being leased, the customer must be constantly vigilant to the danger that the service provider will accidentally misroute the calls through a microwave or satellite channel. Ground routing is often more expensive than satellite routing and might thus be reserved for more sensitive messages. In the past, a similar phenomenon took the form of a choice between a faster route (telephone or telegraph) and a more secure one (physical shipment). If, however, a message has been properly encrypted, the communication network is free to send it any distance, via any channel, without fear of compromise.

One special case of this is particularly important: the freedom to pass an encrypted message through the hands of a potential opponent or competitor. Telecommunications suppliers, for example, often find themselves bidding against the local telephone system to provide equipment or services. In the process of bidding on these contracts, they often have no choice but to employ the services of the communications suppliers with whom they are competing.

Modern security protocols make special use of the ability to pass a message through the hands of a potential enemy. A party to communications is often asked to present cryptographically protected *credentials*. If the credentials can be deciphered correctly, the challenger who receives, deciphers, and judges them need not worry about having received them from a previously untrusted party. This is analogous to the procedure commonly employed to control international travel. When the border guards judge the traveler on the basis of a passport received from the traveler's own hands, they are placing their trust in the tamper resistance of the passport. If the passport appears intact and the picture resembles

the traveler, they will not generally feel the need to conduct any further investigation into the traveler's identity.

In the world of paper documents, this mechanism is not considered adequate for all purposes. A visitor attending a secret briefing at a military installation, for example, must typically be preceded by a letter of authorization. The corresponding cryptographic process, however, is considered reliable enough to be used for the most sensitive applications.

Cryptography can best be thought of as a mechanism for extending the confidentiality and authenticity of one piece of information (the key) to another (the message). The protection of the key, which is typically small and can be handled very carefully, is extended to the message, which can then be handled much less carefully—routed through the least expensive communication channel, for example. As a consequence of this extension of security from the key to the message, the compromise of the key will likewise be extended to the message. The consequences of this compromise, however, differ markedly, depending on whether the compromised key is being used to protect privacy or authenticity.

In regard to authentication, compromise of a cryptographic key is similar to the compromise of other sorts of identifying information, such as passwords or credit cards. Suppose that on Wednesday morning the security officer of a bank learns that an authentication key in use since Monday has been stolen, but also knows that no messages have been received and authenticated with the key since the day before, when the key was known to be safe. The key will be changed immediately and no actual compromise will occur. The thieves cannot make use of a key acquired on Tuesday night to go back and initiate a wire transfer on Monday or Tuesday. If the key is used to protect the privacy of information, however, things are quite different. Even if the compromise of a key is discovered immediately, all messages ever sent in that key must be regarded as compromised. This is because no one can be sure that they were not intercepted and recorded by the same parties who later acquired the key. As soon as the compromise is discovered, the key will of course be changed, but this does far less to repair the damage than the change of an authentication key. The thieves can still read traffic they intercepted while the compromised key was in use. This gives breaches of

cryptographic security the power to reach back in time and compromise the secrecy of messages sent earlier.[31] Changing the key will only prevent future messages from being read.

This vulnerability becomes especially critical in systems, like Alice and Bob's secure telephone in the previous section, that transmit some cryptographic keys encrypted under others. Although session keys in these systems typically last for only the duration of one phone call, session, or transaction, the subscriber keys may stay the same for weeks or months. If one of these is compromised at any time during its life, all the messages ever sent using session keys that were themselves sent enciphered under the subscriber key will likewise be compromised. Worse yet, if the key distribution center is compromised, the messages of every subscriber will be compromised.

Cryptography as we have described it so far has a very centralized character. If keys are distributed physically from a central facility, the center has the power to decide which elements of the network can communicate with which others. If the keys are distributed electronically and in real time, as described for secure phones, the key distribution center acquires control over communication on a virtually call-by-call basis. Implicit in this centralization is the phenomenon that is now called key escrow. The users of a cryptocommunication system may trust it to protect them against external opponents, but they can never be confident that they are protected against the system managers. There is always the possibility that the central facility will be employed to supply additional sets of keys for eavesdropping on the users.

The unique vulnerabilities of conventional cryptography and its centralization are intimately connected. In order to eliminate the former, we must also eliminate the latter.

Public-Key Cryptography

All the cryptographic systems discussed thus far—one-time systems, Vigenère systems, rotor machines, and shift registers—are *symmetric* cryptosystems: the ability to encrypt messages is inseparably linked to the ability to decrypt messages. The course of cryptography for the past 20 years has been dominated by *asymmetric* or *public-key* cryptosystems.

In a public-key cryptosystem, every message is operated on by two keys, one used to encipher the message and other to decipher it (Diffie and Hellman 1976). The keys are inverses in that anything encrypted by one can be decrypted by the other. However, given access to one of these keys, it is computationally infeasible to discover the other one. This makes possible the practice that gives public-key cryptography its name: one of the two keys can be made public without endangering the security of the other. There are two possibilities:

- If the secret key is used for deciphering, anyone with access to the public key will be able to encipher a message so that only the person with the corresponding secret key can decrypt it.
- If the secret key is used for enciphering, its holder can produce a message that anyone with access to the public key can read but only the holder of the secret key could have produced.

The latter property is what characterizes a signed message, and a public-key cryptosystem used in this way is said to provide a *digital signature*.[32]

Using Public-Key Cryptography

In a network using public-key cryptography, the secret key that each subscriber shares with the key management facility in a conventional network is replaced by a pair containing a public key and a private key. The function of the KMF is now merely to hand out public keys. Since these keys are not secret, they are not subject to compromise.

One problem remains, however: a subscriber who receives another subscriber's public key from the KMF must have a way of verifying its authenticity. This problem is solved by providing the KMF with a private signing key and providing every subscriber of the network with a copy of the corresponding public key. This enables the KMF to sign the keys it distributes, and it enables any subscriber to verify the KMF's signature.

Public-key cryptography vastly diminishes the vulnerability of the KMF. If the KMF is compromised, that compromise is the compromise of an authentication key (the KMF's signing key, the only piece of secret information the KMF knows) rather that the compromise of any key used to protect secrecy. If the signing key of the KMF is found to be compromised, that key can be changed. The network's subscribers will have to

be informed of the KMF's new public key and warned not to accept any key distribution messages signed with the old one. Such reinitialization of the network may be costly, but it does not expose past network traffic to disclosure.

This use of public-key cryptosystems also has the practical benefit of reducing the load on the key distribution center. Instead of requiring the subscribers to call the KMF on the occasion of any conversation, they can be provided with a form of credentials called *certificates*. A certificate is a signed and dated message from the KMF to a subscriber containing that subscriber's public key, together with such identifying information as name and address. When two subscribers begin a call, they exchange these credentials. Each one verifies the KMF's signature on the received certificate and extracts the enclosed public key.

Although public-key cryptography as described above does much to diminish the vulnerability of the network as a whole, it does nothing to reduce the vulnerability of individual subscriber's private keys. If a subscriber's private key is compromised, it is possible that any message ever sent in the corresponding public key will be read.

For non-interactive communications, such as electronic mail, there does not appear to be any way around this problem. The person who looks up a public key in a directory and encrypts a message with it is sending the message to the holder of the corresponding private key. Possession of the correct private key is the only thing that distinguishes the receiver, and anyone who has it will be able to read the letter.

For interactive communication, however, the problem can be solved by means of another form of public-key cryptography: *Diffie-Hellman key exchange*,[33] which allows communicating parties to produce a shared secret piece of information despite the fact that all messages they exchange can be intercepted (Diffie and Hellman 1976). Alice produces a secret piece of information that she never reveals to anyone, and derives a corresponding public piece from it. She sends the public piece to Bob and receives from him a piece of public information formed in the same way. By combining their own secret information with the other's public information, Alice and Bob each arrive at a common shared piece of information that they can use as a cryptographic key. The process is analogous to a perfect strategy for bridge bidding in which the North-

South partners (each of whom knows his own hand) agree on a secret but the East-West partners (who do not know either North or South's hand) remain completely ignorant of what North and South have concluded even though they have heard their opponents' every bid and response.

Alice and Bob use the key they have negotiated to encrypt all subsequent messages, but being engaged in encrypted communication is not a sufficient condition for secure communication. Not only does neither Alice nor Bob yet have secure knowledge of the other's identity; they cannot even be sure that an intruder who has performed a key exchange with each of them is not sitting in the middle, translating between keys and recording the plain text of their conversation.

The second step, therefore, is to exchange certificates: Alice sends Bob hers, and he sends her his. Verifying the KMF's signature on the received certificates allows the receiver to be sure that the certificate is authentic but does not guarantee the authenticity of the person who sent the certificate. It remains, therefore, for Alice (for example) to verify that the person with whom she is in contact is actually the legitimate holder of Bob's certificate—the person who knows the secret key corresponding to the public one it contains. This final step is done by a process called *challenge and response* in which the challenger sends a test message and judges the legitimacy of the responder by verifying the signature on the response. In this case, it is best to use as the challenge the piece of public information from the exponential key exchange, since so doing gives assurance that the encryption is actually being performed by the same entity that engaged in the authentication process.

Packet communication, on which the Internet is based, has made complex cryptographic protocols vastly more feasible. In a call-based telephone system, the only way of providing a common resource is to make it available to be called—as an 800 number, for example. This suffers from two problems. Typical use of a common resource in cryptography is brief: call, send and receive a few hundred to a few thousand bits (usually a key or certificate), and hang up. Such a call is inherently expensive because it uses the same circuit setup machinery as a call of much greater length.[34] Second, unless the phones have conference calling capability, they can only make one call at a time, so the protocol must be organized into a sequence of non-overlapping connections for each party.[35]

Packet switching functions much like the mail but at the speed of the telephone. A packet is a small collection of data: an address, a return address, a size field, the contents, and some flags. Typical packets vary from tens to thousands of bytes in length; tiny in networks operating at thousands to millions of bytes a second. Energetic correspondents might send each other hundreds or even thousands of pages in the course of a year; similarly, devices on the Internet are free to send arbitrarily large amounts of data, a little bit at a time in the "small" packets. The analogy to a postal correspondence is a connection. Like a postal correspondence, the connection uses the resources of the communication system only when a packet (letter) is in transit. At other times, it consumes nothing but a line in an address book.

The protocols that implement communication on the Internet are organized into layers. The lower layers deal with characteristics of the physical communication system and have the suggestive names *physical* and *link*. The topmost layer is called *application* and is inhabited, as its name suggests, by the computer programs doing the communicating. In the middle are two critical layers, whose names are almost as instructive, called *network* and *transport*.

The network layer is all important; it is what defines the network. For two nodes to be on the same network they must have exactly the same concept of addressing. It is the Internet Protocol (IP) that specifies how Internet addressing works.[36] The Internet Protocol provides unreliable delivery packet delivery; it does its best to get each packet to its destination in a timely fashion. If the time runs out or if a packet is lost to equipment failure, the packet is lost. The packet will generally be retransmitted but this is the responsibility of protocols in the layer above.

The transport layer is the layer that accommodates itself to the varying characteristics of Internet traffic. The most frequently used transport-layer protocols is called the Transmission Control Protocol and provides reliable communication—through error detection codes and retransmission—on top of the unreliable service provided by the layer below.

Packet-switched systems are free from both of circuit switching's problems. One-time transmission of small numbers of bits is inexpensive and a device connected to a packet-switched system can readily be in contact with many devices at the same time. This has given rise to a variety of

packet-network security protocols, including SSL and the Internet Protocol Security Protocol (IPSec).

Because the network layer is shared across the network, it is the obvious place to install cryptographic protection. The operation of the network layer is not entirely compatible with cryptography, however. Each packet carried by IP is independent of every other packet, even between the source and destination. Encryption of traffic between two points typically uses the same key for substantial periods of time. A satisfactory compromise between these factors took some time to achieve. IPsec standardization began to take shape in the late 1990s.[37]

A more natural place to put security is in the transport layer because this layer already has the facilities for associating packets and keeping track of information flowing in streams. The best-known and most widely used security mechanism on the Internet is SSL. This facility is embodied in all browsers and implements https, the secure version of the hypertext transport protocol that delivers web pages.

Communication Security

The indispensable application of cryptography is the protection of communication. How this is accomplished depends on the form of the communication network and on whether the protection is being applied by the network's owners, by the subscribers, or by communication providers who supply the network with particular communication resources (such as satellite channels).

There are three basic ways in which encryption can be applied to network communication:

- *Net keying*. Every element in the network uses the same key, which is changed at regular intervals (often daily). This is the key-management technique employed by the US Navy's Fleet Broadcast System and many other government networks.
- *Link keying*. Messages are encrypted as they go from switching centers onto communication channels and decrypted as they go back into switching centers. A message is thus encrypted and decrypted several times as it goes from sender to receiver. This tech-

nique permits addressing information as well as message contents to be encrypted and makes the traffic particularly inscrutable to anyone outside the network.

- *End-to-end keying*. Each message is encrypted as it leaves the sender and cannot be decrypted by anyone other than the receiver. This is the typical behavior of secure telephones and secure email.

The results achieved by applying cryptography to secure communication networks depend dramatically on where it is applied. The most effective way for two individuals or two organizations to communicate securely over a network they do not control is to encrypt their communications using end-to-end keying. Relying on measures applied by network management may protect them from most opponents but will always leave them vulnerable to a foe with the resources to seek out a weak point along their communication path and exploit it. On the other hand, network keying and link keying allow the network to provide the users with security services they cannot provide for themselves.

Transmission Security and Covert Communication

Through a technique known as *traffic analysis*—the study of the patterns of communication—an opponent can learn a great deal about the activities of an organization without being able to understand any individual message. The counter to traffic analysis, *transmission security* or *communications cover*—which always amounts to sending dummy traffic to conceal the pattern of real traffic (Kent 1977)—is difficult and expensive to implement on an end-to-end basis and is best left to the network infrastructure.

The most extreme form of transmission security is to conceal the existence of communication altogether. This is often done by using *frequency-hopping radios*, whose frequencies change many times a second in an unpredictable pattern. When concealment of the existence of communication is the primary objective, we speak of *covert communication*. This is the dominant concern in the communications of criminals and spies.

Supporting Technologies

In order to be effective in network security, cryptography must not only be employed for the right purposes, it must also be implemented correctly. Several *supporting technologies* play important roles in this respect.

Reliability is critical in secure communication. Failures of either cryptographic equipment or its human operators, called *busts*, are typically the most lucrative sources of cryptanalytic successes (Welchman 1982). Automation has made a major contribution to the ease of use of cryptoequipment, making human errors less likely, but rising data rates have made reliability of the equipment vastly more significant. Performing a *security failure analysis* to trace the effects of all likely failures and including self-check and failure-monitoring circuitry in essential in designing cryptographic equipment.

Secure computing practices are integral to the construction of reliable communication-security (COMSEC) equipment whose functioning can be trusted under adverse conditions. A particularly difficult aspect of the logical analysis is the detection of malicious hidden functions, and no good solution to this problem is known for equipment acquired from untrusted suppliers.

Electromagnetic shielding is essential to prevent cryptographic channels from being bypassed by radiative or conductive emissions or by accidental modulation of a ciphertext signal with a plaintext signal. The military term for this particular form of protection is Tempest.[38]

The most interesting and difficult Tempest problem is the contamination of ciphertext by plaintext. In the United States, this problem first seems to have been observed in the late 1940s or the early 1950s in an online version of a one-time-tape system called SIGTOT (Martin 1980, pp. 74–75). What they observed was probably a form of amplitude modulation in which combining a 0 in the key with a 1 in the plaintext produces a waveform that is distinguishable from the waveform produced by combining a 1 in the key with a 0 in the plaintext, even though the results are supposed to be identical. If this occurs, opponents with the right equipment can read both the plaintext and the keystream from the ciphertext signal. The effect as described is unlikely to occur in modern digital cryptosystems, but amplitude modulation is only the sim-

plest form of plaintext contamination. Frequency and timing modulation present subtler pitfalls for the designer.

Worried as people tend to be about surprises in mathematics undermining cryptosystems, in recent years implementation has proven the richest source of out-of-the-box thinking leading to surprising compromises of cryptographic equipment. No attack exemplifies this better than Paul Kocher's *differential power analysis*. Kocher showed that by measuring the the varying power demands of a microprocessor or a dedicated cryptographic chip. This problem can be overcome in "large" pieces of cryptographic equipment that contain batteries and power supplies and present a constant demand for recharging power to the outside world. Smart cards, however, have negligible power storage and depend on a constant supply of external power and countering attacks of this kind in this environment is extremely difficult. Kocher's work has subsequently been generalized and many cases have been found in which a measurable external symptom has been found to correlate with a cryptographically significant event in algorithm execution. Often it is just a question of whether the algorithm is processing a key bit that is 0 or a key bit that is 1. In such cases, the key can be read out directly, with disastrous results for security.

Tamper resistance guarantees that COMSEC equipment will not be altered to defeat its functioning without requiring the expensive practice of guarding it constantly or locking it in safes when not in use. Tamper resistance dramatically reduces the level of trust that must be placed in personnel permitted to operate COMSEC equipment and has become a mainstay of US military cryptography.

The Operation of Secure Communications

A secure communication system must not only be designed and built correctly; it must be operated correctly over its entire lifetime. The threat posed by opponents must be assessed, not once when a system is developed, but continually. The vulnerability must be judged against advancing technology. This should include continuing cryptanalytic study of the cryptosystems, other sorts of penetration studies, and continuing reevaluation of who the opponents are and what their resources are.

Equipment that worked securely when it was new may not continue

to work securely. Ongoing assessment of the functioning of installed equipment must be accompanied by a careful program of maintenance.

A vital element of a secure communication posture that is operational rather than architectural is *communication security monitoring*. This is the practice of intercepting friendly communications to monitor the effectiveness of COMSEC practices. COMSEC monitoring is a difficult and sensitive task that is rarely undertaken by non-governmental organizations.

Cryptographic Needs of Business

Many large American firms have manufacturing plants around the world and need to communicate product-design information, marketing plans, bidding data, costs and prices of parts and services, strategic plans, and orders to them. Much of this information, which is often communicated electronically, must be kept secret from competitors.[39]

In the current banking system, the transfer of currency is the transfer of electronic bits, and it could not be undertaken without adequate security. Cryptography is simply the latest manifestation of the security upon which the banking industry has always relied. Internationalization of banking complicates the security problem while exacerbating the need for security.

Except to the extent that loss of confidentiality threatens security by exposing access controls, authentication is more important in banking than data confidentiality.[40] In part, this is because exposure of one person's data typically harms only that individual, whereas failures of authentication can result in loss of assets. In part, it is because there is already substantial government monitoring of financial transactions.[41]

Electronic funds transfer is already several decades old, but bankers now face the complex security issues posed by opening the transfer mechanisms to a wider audience. Indeed, despite stringent security procedures, Citicorp has already been the victim of such a scam, losing $400,000 in a theft that occurred over several months and involved three continents.[42] After this electronic theft became public, six of Citicorp's competitors went after its largest accounts, claiming they could provide better security than Citicorp (Carley 1995; Hansell 1995).

Although banking is a highly regulated industry, that does not mean that the government and banks see eye-to-eye on cryptography. In the 1980s, the National Security Agency proposed replacing the Data Encryption Standard with a classified algorithm. Bankers, having invested heavily in designing protocols based on DES, protested, and the NSA plan was dropped. Bankers also objected to the Clipper Chip.

In the oil industry seismic data and other geographic information can be worth a small fortune. Even before they came to rely on electronic communications, oil companies were targets of electronic eavesdropping.[43] Oil companies often operate in politically unstable regions, and criminal actions against employees, including kidnappings, are not unknown. Thus, in order to minimize knowledge about their whereabouts, employees' communications must be secured (Dam and Lin 1996, p. 464).

Other businesses face different requirements. In contrast to banking, the health-care industry emphasizes confidentiality. Medicine's heavily integrated systems of insurance records, hospital files, and doctors' records require a system that preserves patient confidentiality while allowing access by a large set of users (NRC 1997). The Health Insurance Portability and Accountability Act (HIPAA) has become one of the major drivers of privacy technology in the commercial world, and protection of personal data leans heavily on encryption.

Knowledge-based industries seek to protect their heavy investments in intangible goods, and cryptography will be central to that protection.

The entertainment industry has sought to protect digitized content through a combination of legislation[44] and technology. It has embraced cryptography as the way to make first-run movies and other expensive products available online while retaining close control that prevents the viewers from making copies or making any use of the products other than those the providers approve. The essence of this technology, called *Digital Rights Management* (DRM), is to keep information in encrypted form everywhere except inside tamper-resistant devices. In Hollywood's vision, a movie would travel encrypted, perhaps all they way from initial production at the studio until it reached the processor in the screen of your TV set.

A major initiative called *Trusted Platform Technology*, which seeks

to install tamper-resistant security hardware into computers, particularly personal computers, laptops, and smaller devices, has been under way for nearly a decade. This technology will be of inestimable value in securing national critical infrastructure but threatens to diminish the degree of control users can exercise over their own machines.

Although digital digital signatures can provide a guarantee that movies, recordings, and other works of art have not been modified, a digital signature can be removed without trace. In addition to direct attempts to control the use of digital materials, content providers are also using cryptography to embed *watermarks* that permit individual copies to be identified and tracked. A watermark is a sort of tamper-resistant digital signature that makes the origin of digital information difficult to conceal.

Economic espionage is not solely the province of competing companies; it is also widely practiced by governments. One of the top priorities of the French intelligence service is industrial spying. Pierre Marion, a former director of the Direction Générale de la Sécurité Extérieure, the French Intelligence Service, told NBC News that he had initiated an espionage program against US businessmen to keep France economically competitive.[45] Another French spy, Henri de Marenches, the director of the French secret service from 1970 to 1981, observed that economic spying was very profitable: "It enables the Intelligence Services to discover a process used in another country, which might have taken years and possibly millions of francs to invent or perfect." (Schweizer 1993, p. 13) The British wiretap law includes protecting the economic well-being of the country as one of the reasons to permit wiretapping.[46] The Japanese invest heavily in industrial espionage, sending hundreds of businessmen abroad to discover in detail what the competition is doing, hiring many foreign consultants to further the contacts, and electronically eavesdropping on foreign businessmen in Japan (ibid., pp. 74–82 and 84). In China, the office of a multinational company experienced a theft in which unencrypted computer files were copied (ibid., p. 471).

In many countries, telecommunications are run by the national government. This makes state electronic eavesdropping particularly simple. The governments of Japan and France are notorious for eavesdropping on the communications of US businessmen (Schweizer 1993, pp. 16 and 84). It is alleged that the Bundesnachrichtendienst, the German intelligence

service, regularly wiretaps transatlantic business communications (ibid., p. 17).

Similar charges have been raised against the US government. There was suspicion in Britain for some time that the NSA establishment at Menwith Hill was being used as much to spy on the British as on the Eastern Europeans (Bamford 1982, p. 332). These suspicions became more general in 2000 with the exposure of a US intelligence network called *ECHELON* that was largely devoted to monitoring commercial communication channels around the world. (Campbell 1999) From one viewpoint this was a natural outgrowth of the growing use of commercial rather than purpose-built communication systems by military organizations around the world, but this did little to soothe the worries of those who felt they were being spied on.

In 1990 FBI Director William Sessions said that the FBI would devote greater investigative energies to the intelligence activities of "friendly" nations (Schweizer 1993, p. 4). Of course, greater protection of the goods to begin with would decrease the need to investigate thefts after they had occurred—a point that is surely not lost on the FBI.

In another sort of case, an American businessman working for a multinational firm reported that his laptop computer was taken by customs officials of an unnamed country and returned to him three days later; as he attempted to negotiate his business deals, it became clear that his sensitive files had been copied during those three days (Dam and Lin 1996, p. 471). With the widespread use of laptops and other portable devices (AP 2006) by large numbers of workers not necessarily versed in information security, the loss of such devices and the data they contain has become an almost daily news item. The potential compromise of the sensitive data they contain has led to a federal government requirement that all data on laptops used by federal agencies be encrypted unless the data are determined to be nonsensitive by a designee of the Deputy Secretary of the agency (Johnson 2006).[47]

It has been claimed that a system like public key, in which the private key is stored with a single user, will not provide the data-recovery features that corporations require. In fact, the danger posed to corporations by this lack of data recovery for communications is minimal. With the exception of the financial industry, few businesses currently record tele-

phone communications. Those that do can continue to do so even if the telecommunications are encrypted. Because the ends of the conversation will be unencrypted, and the recording can be done at that point, the choice of encryption—whether public key or some form of escrow—will not affect data recovery. Encrypted communications will provide corporations with at least as much security as they have now; they will not be losing information, for *they do not have such information in the first place*.

Because they rely on "written" records rather than recordings of conversations, businesses are in the same position as law-enforcement agencies. Conversations are transient whereas records endure, and wiretaps are used in far fewer criminal cases than seized records (Dam and Lin 1996, p. 84).

In the mid 1990s, talk of a global information infrastructure, an information superhighway, and Internet commerce was everywhere. The sense was that we were moving our culture into digital channels of communication. A decade later, the Internet has indeed transformed the life of the industrialized world. Email is replacing mail; websites are replacing catalogs and advertising; weblogs and search engines are replacing newspapers. Governments, corporations, and terrorist organizations make contact with their constituencies over the World Wide Web.

In the first edition of this book, we said of the move to the digital world:

> If this is to be a success, we will to have to find digital replacements for a lot of everyday physical practices. In the area of security, many of the new practices will be cryptographic.

In some cases, the correspondences are obvious. In the physical world, we close doors or stroll off somewhere by ourselves or whisper in order to have privacy. In the digital world, we encrypt. In other cases, the correspondence is not so obvious. In the physical world, we place written signatures on contracts, letters and checks. In the digital world, we add digital signatures and rejoice in the degree of similarity. Some of the correspondences are still unresolved and are controversial. A copying machine may reproduce a readable copy of a bestseller, but for most purposes the copy is a poor imitation of the original. In the digital world, if you can get your hands on a document you can copy it exactly. A system

that prevents the unauthorized copying of a digital novel, however, is capable of preventing the reader from making many legitimate uses of it. Many of these issue remain unresolved.

So where do we stand? On one hand, the Secure Socket Layer protocol that underlies secure browsing (https) is perhaps the most widely deployed cryptographic system in the world. On the other, the news abounds with stories of spam, breakins, phishing, and identity theft. It is reasonable to conclude that not only will a lot of everyday security practices have to find digital replacements but a lot of new ones will have to be developed.

Digital equivalents have been found for a surprising range of human interactions, including:

- The delivery of a registered letter. (You get the letter if and only if the deliverer gets a signature on the receipt.)
- The signing of a contract. (The contract is valid only if both signatures are on both copies and neither party can get a final copy while denying one to the other.)
- Sharing of authority. (The president of the company can sign a million-dollar check, but it takes two vice presidents.)
- Power of attorney. (My lawyers have access to the contents of my safe deposit box because they are working for me.)
- Issuing of credentials. (A passport certifies that the secretary of state vouches for the bearer.)

Why Has Cryptography Taken So Long to Become a Business Success?

If cryptography is so valuable, why has it taken so long to become a business success? One answer lies in the truly remarkable qualities of communications intelligence. Unlike installing an infiltrator, breaking and entering, or going around asking questions, interception of communications is very hard to detect. People may suspect that their communications are being spied on, but they are rarely able to prove it or to convince themselves with certainty that that is happening. Even an intelligence

agency typically has difficulty discovering that it is being spied on and how. During the 1960s and the 1970s, the British and American intelligence establishments were convinced that they had been penetrated by the Soviets and spent an excessive amount of their time trying to discover who the turncoats were (Wright 1987). Despite their efforts, the investigations were inconclusive.

For an organization that is not an intelligence agency, such *counterintelligence* is much harder. Merely discovering that information has been leaking is very difficult. Discovering how or to whom is far more difficult. It may even require an intelligence agency to uncover the fact of the eavesdropping; this has happened several times for American companies.[48] Selling secure communications has often been likened to selling insurance, in that the customer must pay up front for protection against something that may occur in the future. The fire, auto, and medical insurance salesmen have an advantage, however. Everyone has seen houses burn down, cars crash, and friends get seriously ill. Almost no one has seen information taken by eavesdroppers and used to bankrupt an otherwise profitable business. Indeed, admitting to break-ins has its own costs: competitors may use the seeming lack of security to woo customers.

The Problem of Standards

Cryptography also suffers from a serious standards problem. In a sense, security equipment exists to amplify minor incompatibilities into absolute non-interoperability. If a radio is not tuned to quite the right frequency, reception will suffer but communication may still be possible. If a cryptographic key is off by even one bit, communication is impossible. By themselves standards problems probably would not account for cryptography's slow start, but in combination with other factors they have played a major role. In particular, the lack of standards has contributed directly to the lack of "critical mass." A single secure telephone, just like a single telephone, is a useless piece of hardware, and even a pair has only limited applicability. Only when there is a proliferation of such devices does their value increase to the point that purchasing a security device for one's telephone is as common as purchasing a lock when buying a bicycle.

Similarly, the lack of a supporting infrastructure has slowed the adop-

tion of secure communications. Without world-wide keying infrastructure and key-management facilities, the provision of keys is a remarkably cumbersome process, and encrypted communications are tend to be used only by thosec committed to security.

The Adverse Effect of Government Opposition

In the United States, cryptography was long hurt by the ambivalent attitude of the federal government. Despite the controversy that surrounded the adoption of the Data Encryption Standard in 1977, this standard and the others that accompanied it have made a major contribution to the deployment of cryptography both in the US and abroad. This latter aspect made the government wonder whether it had done the right thing, and in the 1980s it became an antagonist rather than a proponent of cryptographic standards.

The delays caused by the 1990s crypto wars are part of the reason for the poor state of laptop security. Now the government finds itself in the peculiar situation of insisting on ubiquitous security—and cryptography—on consumer laptops, at least those purchased by the federal government.

The Effect of the Internet

Since 1995, the market for cryptography has exploded. What has changed? Most conspicuously, the Internet.

The Internet has made global electronic commerce a day-to-day reality for a large number of people.[49] And commerce, on a large scale, can prosper only when people can deal confidently with people they have never met and have no reason to trust. The problem is made worse by the Internet's internationality. No uniform system of law or policing can patrol it. The merchants, like the cannon-carrying merchant ships of two centuries ago, must provide security themselves. The more secure people can be in their transactions, the larger those transactions will be and the more profitable the Internet will be as a business medium.

At present, the security of Internet commerce must be judged as fair to middling. The tension between strong identification and anonymity has yet to be resolved and a uniform public-key infrastructure has yet to be built. Nonetheless, the Internet now accounts for a significant fraction of most businesses' business. It has created new businesses—eBay's world-

wide auctioning of low-value items, Google—and transformed others. Some used book stores have been driven out of business while others have thriven by putting their wares online.

The last decade saw substantial societal change in the usage of mobile communications technologies: cell phones, PDAs, etc. Communication patterns changed, not just among the young with cellphones and SMS, but also among the traveling public, and, at least in certain societies, an expectation has arisen that communication devices are always on and people are always available.

As bandwidths rise and computing technology advances, telecommunications increasingly have become preferred modes of communication rather than merely increasingly satisfactory substitutes for pre-electronic modes.

There has been a simultaneous explosion of online communications using the Internet. The traditional communication tools—if anything less than a century old can actually be called traditional—are email, which in less than a decade has become ubiquitous,[50] instant messaging, and voice calls using Internet technology (VoIP, or Voice over IP). There are some quite non-traditional communication methods as well, such as music-jamming sessions over the Internet and the very popular *massively multiplayer on-line role playing games* (*MMORPG*s).

Although MMORPGs are called games, they are a recreation with as much in common with hanging out in the mall as with the videogames you might find there. Some MMORPGs are more focused on the prize (or on reaching "levels") than others, but MMORPGs differ from console games exactly because they enable interaction with other players. (Until the Internet was opened to commercial use, MMORPGs were limited to online services such as AOL and CompuServe, and all players in a particular game had to subscribe to that service. Once that changed, MMORPGs began to appear on the public Internet.) MMORPGs provide an online environment in which players, through their avatars (icons representing the user), slay dragons, compete for prizes, communicate, and buy, trade, and sell fictional objects of value. The fictional objects, moreover, are no longer entirely fictional. Often they are bought and sold on eBay and in other online venues.

MMORPGs are communities and yet another venue for electronic communications, one far from traditional, but widely used, especially

in Asia. In the United States, roughly 19 percent of the gaming population are online gamers, but in the Asia/Pacific region, consisting of Hong Kong, Korea, Malaysia, the People's Republic of China, Singapore, and Taiwan, computer games (which means online games, as console games are virtually non-existent in the region) are extremely popular. There are two reasons driving the interest in Asia: in China, Internet cafes are open around the clock, and Korea is experiencing rapid broadband adoption (IDC 2006).

What Is Cryptography's Commercial Future?

In light of the sudden growth of cryptography, it is natural to ask how big a business it will become. In order to find the answer, we might look at an existing business.

Locksmithing is not a large fraction of the building trades. If you own a $500,000 house you are unlikely to have $500 worth of locks on it. Even if, like Oliver North, you have a fancy electronic security system, its total value probably doesn't exceed a few percent of the house's value. On the other hand, no one would argue that locks are not essential to the functioning of society. The modern suburb, in which most members of most families are gone for most of the day, would not function without them.

A harbinger in the cryptographic world is Skype, a Voice over IP communication system that encrypts all of its transmissions over the Internet automatically. Security isn't the point of Skype, as it is with Zfone[51]; it is just part of the package. Skype was built by a group of Latvians, now in their mid thirties, who grew up under Soviet rule and who take it for granted that someone is trying to listen in on their phone calls. One of the most popular Voice over IP systems, Skype, encrypts the voice traffic in every call without any explicit action by the user.

Cryptography seems to be most successful where it is following a similar course, not a prominent part of any product but ever present and essential. The Secure Socket Layer protocol in Web browsers is a candidate for being the most widely deployed cryptographic protocol of all time. It could be overtaken by the automatic encryption in Skype. Skype is also an example of the internationalization of cryptotechnology: the product originated in Latvia but is now owned by eBay. It may also be the focus of the US government's latest policy efforts to capture communications traffic, an issue we discuss in chapter 11.