FIGURE 1.15 Solution of substitution cipher.

Key	Message
0:	L Z W J W A K F G G L Z W J D S F Y M S Y W T M L X J W F U Z
1:	K Y V I V Z J E F F K Y V I C R E X L R X V S L K W I V E T Y
2:	J X U H U Y I D E E J X U H B Q D W K Q W U R K J V H U D S X
3:	I W T G T X H C D D I W T G A P C V J P V T Q J I U G T C R W
4:	H V S F S W G B C C H V S F Z O B U I O U S P I H T F S B Q V
5:	G U R E R V F A B B G U R E Y N A T H N T R O H G S E R A P U
6:	F T Q D Q U E Z A A F T Q D X M Z S G M S Q N G F R D Q Z O T
7:	E S P C P T D Y Z Z E S P C W L Y R F L R P M F E Q C P Y N S
8:	D R O B O S C X Y Y D R O B V K X Q E K Q O L E D P B O X M R
9:	C Q N A N R B W X X C Q N A U J W P D J P N K D C O A N W L Q
10:	B P M Z M Q A V W W B P M Z T I V O C I O M J C B N Z M V K P
11:	A O L Y L P Z U V V A O L Y S H U N B H N L I B A M Y L U J O
12:	Z N K X K O Y T U U Z N K X R G T M A G M K H A Z L X K T I N
13:	Y M J W J N X S T T Y M J W Q F S L Z F L J G Z Y K W J S H M
14:	X L I V I M W R S S X L I V P E R K Y E K I F Y X J V I R G L
15:	W K H U H L V Q R R W K H U O D Q J X D J H E X W I U H Q F K
16:	V J G T G K U P Q Q V J G T N C P I W C I G D W V H T G P E J
17:	U I F S F J T O P P U I F S M B O H V B H F C V U G S F O D I
18:	T H E R E I S N O O T H E R L A N G U A G E B U T F R E N C H
19:	S G D Q D H R M N N S G D Q K Z M F T Z F D A T S E Q D M B G
20:	R F C P C G Q L M M R F C P J Y L E S Y E C Z S R D P C L A F
21:	Q E B O B F P K L L Q E B O I X K D R X D B Y R Q C O B K Z E
22:	P D A N A E O J K K P D A N H W J C Q W C A X Q P B N A J Y D
23:	O C Z M Z D N I J J O C Z M G V I B P V B Z W P O A M Z I X C
24:	N B Y L Y C M H I I N B Y L F U H A O U A Y V O N Z L Y H W B
25:	M A X K X B L G H H M A X K E T G Z N T Z X U N M Y K X G V A

M = THERE IS NO OTHER LANGUAGE BUT FRENCH.†

Because only one of the keys (K = 18) produces a meaningful message, we have:

$$p_C(M) = 1$$
$$p_C(M') = 0, \text{ for every other message } M'$$
$$p_M(C) = p(18) = \frac{1}{26}$$
$$p_{M'}(C) = 0, \text{ for every other message } M'. \blacksquare$$

Example:
With a slight modification to the preceding scheme, we can create a cipher having perfect secrecy. The trick is to shift each letter by a random amount. Specifically, K is given by a stream $k_1 k_2 \ldots$, where each k_i is a random integer in the range [0, 25] giving the amount of shift for the ith letter. Then the 31-character ciphertext C in the preceding example could correspond to any valid 31-character message, because each possible plaintext message is derived by some key stream. For example, the plaintext message

THIS SPECIES HAS ALWAYS BEEN EXTINCT.†

is derived by the key stream

18, 18, 14, 17, 4, ...

Though most of the 31-character possible plaintext messages can be ruled out as not being valid English, this much is known even without the ciphertext. Perfect secrecy is achieved because interception of the ciphertext does not reveal anything new about the plaintext message.

The key stream must not repeat or be used to encipher another message. Otherwise, it may be possible to break the cipher by correlating two ciphertexts enciphered under the same portion of the stream (see Section 2.4.4). ■

A cipher using a nonrepeating random key stream such as the one described in the preceding example is called a **one-time pad**. One-time pads are the only ciphers that achieve perfect secrecy. Implementation of one-time pads and approximations to one-time pads is studied in Chapters 2 and 3.

1.4.3 Unicity Distance

Shannon measured the secrecy of a cipher in terms of the **key equivocation** $H_C(K)$ of a key K for a given ciphertext C; that is, the amount of uncertainty in K given C. From Eq. (1.2b), this is

$$H_C(K) = \sum_C p(C) \sum_K p_C(K) \log_2 \left(\frac{1}{p_C(K)}\right) ,$$

where $p_C(K)$ is the probability of K given C. If $H_C(K)$ is 0, then there is no uncertainty, and the cipher is theoretically breakable given enough resources. As the length N of the ciphertext increases, the equivocation usually decreases.

The **unicity distance** is the smallest N such that $H_C(K)$ is close to 0; that is, it is the amount of ciphertext needed to uniquely determine the key. A cipher is **unconditionally secure** if $H_C(K)$ never approaches 0 even for large N; that is, no matter how much ciphertext is intercepted, the key cannot be determined. (Shannon used the term "ideal secrecy" to describe systems that did not achieve perfect secrecy, but were nonetheless unbreakable because they did not give enough information to determine the key.)

Most ciphers are too complex to determine the probabilities required to derive the unicity distance. Shannon showed, however, it is possible to approxi-

† From S. Gorn's Compendium of Rarely Used Cliches.

† Also from S. Gorn's Compendium of Rarely Used Cliches.

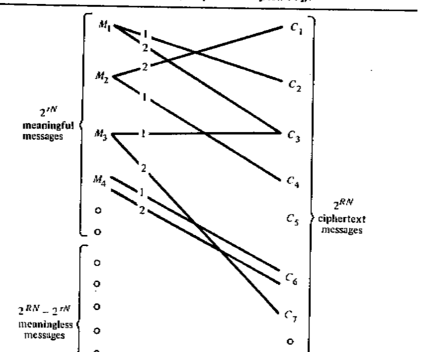
rived Shannon's result using a slightly different approach.

Following Hellman, we assume each plaintext and ciphertext message comes om a finite alphabet of L symbols. Thus there are 2^{RN} possible messages of ngth N, where $R = \log_2 L$ is the absolute rate of the language. The 2^{RN} messages e partitioned into two subsets: a set of 2^{rN} **meaningful messages** and a set of 2^{RN} 2^{rN} **meaningless messages**, where r is the rate of the language. All meaningful essages are assumed to have the same prior probability $1/2^{rN} = 2^{-rN}$, while all eaningless messages are assumed to have probability 0.

We also assume there are $2^{H(K)}$ keys, all equally likely, where $H(K)$ is the key tropy (number of bits in the key). The prior probability of all keys is $p(K)$ $1/2^{H(K)} = 2^{-H(K)}$.

A **random cipher** is one in which for each key K and ciphertext C, the cipherment $D_K(C)$ is an independent random variable uniformly distributed over 2^{RN} messages, both meaningful and not. Intuitively, this means that for a given and C, $D_K(C)$ is as likely to produce one plaintext message as any other. Actual- the decipherments are not completely independent because a given key must iquely encipher a given message, whence $D_K(C) \neq D_K(C')$ for $C \neq C'$.

FIGURE 1.16 Random cipher model (adapted from [Hell 77]).



the ciphertext $C = E_K(M)$ for given K and M. A **spurious key decipherment** or **false solution** arises whenever encipherment under another key K' could produce C; that is, $C = E_{K'}(M)$ for the same message M, or $C = E_{K'}(M')$ for another meaningful message M'. Figure 1.16 shows two spurious key decipherments, one from the third ciphertext and one from the sixth. A cryptanalyst intercepting one of these ciphertexts would be unable to break the cipher since there would be no way of picking the correct key. We are not concerned with decipherments that produce meaningless messages, because the cryptanalyst can immediately reject these solutions.

Now, for every correct solution to a particular ciphertext, there are $(2^{H(K)} - 1)$ remaining keys, each of which has the same probability q of yielding a spurious key decipherment. Because each plaintext message is equally likely, the probability of getting a meaningful message and, therefore, a false solution is given by

$$q = \frac{2^{rN}}{2^{RN}} = 2^{(r-R)N} = 2^{-DN},$$

where $D = R - r$ is the redundancy of the language. Letting F denote the expected number of false solutions, we have

$$F = (2^{H(K)} - 1)\, q = (2^{H(K)} - 1)\, 2^{-DN} \simeq 2^{H(K)-DN}. \tag{1.3}$$

Because of the rapid decrease in the exponential with increasing N,

$$\log_2 F = H(K) - DN = 0$$

is taken as the point where the number of false solutions is sufficiently small the cipher can be broken. Thus

$$N = \frac{H(K)}{D} \tag{1.4}$$

is the unicity distance—the amount of text necessary to break the cipher.

If for given N, the number of possible keys is as large as the number of meaningful messages, then $H(K) = \log_2(2^{RN}) = RN$; thus

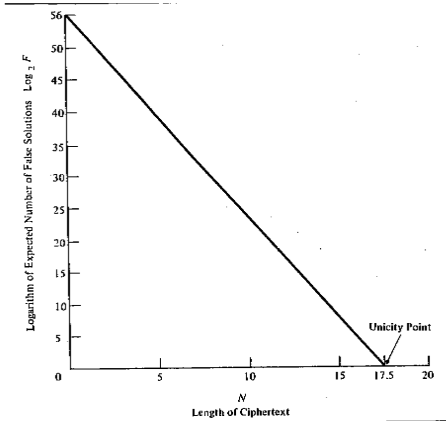
$$H(K) - DN = (R - D)N = rN \neq 0,$$

and the cipher is theoretically unbreakable. This is the principle behind the one-time pad.

Example:
Consider the DES, which enciphers 64-bit blocks (8 characters) using 56-bit keys. The DES is a reasonably close approximation to the random cipher model. Figure 1.17 shows F as a function of N for English language messages, where $H(K) = 56$ and $D = 3.2$ in Eq. (1.3). The unicity distance is thus

$$N = \frac{56}{3.2} = 17.5 \text{ characters},$$

or a little over two blocks. Doubling the key size to 112 bits would double the unicity distance to 35 characters. ■



Example:
Consider a simple substitution cipher that shifts every letter in the alphabet forward by K positions, $0 \leq K \leq 25$. Then $H(K) = \log_2 26 = 4.7$ and the unicity distance is

$$N = \frac{4.7}{3.2} = 1.5 \text{ characters}.$$

This estimate does not seem plausible, however, because no substitution cipher can be solved with just one or two characters of ciphertext. There are two problems with the approximation. First, the estimate $D = 3.2$ applies only to reasonable long messages. Second, the cipher is a poor approximation to the random cipher model. This is because most ciphertexts are not produced by meaningful messages (e.g., the ciphertext QQQQ is produced only by the meaningless messages AAAA, BBBB, . . . , ZZZZ), whence the decipherments are not uniformly distributed over the entire message space. Nevertheless, shifted ciphers can generally be solved with just a few characters of ciphertext. ■

ciphertext needed to break a cipher. Thus a particular cipher will have a unicity distance of at least $H(K)/D$. In practice, $H(K)/D$ is a good approximation even for simple ciphers. We shall derive the unicity distance of several ciphers in Chapter 2. The interested reader can read more about the unicity distances of classical ciphers in Deavours [Deav77].

The unicity distance gives the number of characters required to uniquely determine the key; it does not indicate the computational difficulty of finding it. A cipher may be computationally infeasible to break even if it is theoretically possible with a relatively small amount of ciphertext. Public-key systems, for example, can be theoretically broken without any ciphertext at all. The cryptanalyst, knowing the public key and the method of generating key pairs, can systematically try all possible private keys until the matching key is found (see Brassard [Bras79a, Bras80]). This strategy is computationally infeasible, however, for large key spaces (e.g., with 2^{200} keys). The DES can also be broken by exhaustive search of the key space in a known-plaintext attack (by trying all keys until one is found that enciphers the plaintext into the matching ciphertext). Nevertheless, the best known strategies for breaking the DES are extremely time-consuming. By contrast, certain substitution ciphers discussed in the next chapter use longer keys and have much greater unicity distances than DES. These ciphers are often relatively simple to solve, however, when enough ciphertext is intercepted.

Equation (1.4) shows that the unicity distance N is inversely proportional to the redundancy D. As D approaches 0, an otherwise trivial cipher becomes unbreakable. To illustrate, suppose a 6-digit integer M is enciphered as 351972 using a Caesar-type shifted substitution cipher with key K, where $0 \leq K \leq 9$, and that all possible 6-digit integers are equally likely. Then a cryptanalyst cannot determine which of the following integers is the value of M:

Key	*Integer*
0	351972
1	240861
2	139750
.	.
.	.
.	.
9	462083

The reason the cipher cannot be solved is that the language has no redundancy; every digit counts.

Because of the inherent redundancy of natural languages, many ciphers can be solved by statistical analysis of the ciphertext. These techniques use frequency distributions of letters and sequences of letters, ciphertext repetitions, and probable words. Although a full discussion of these techniques is beyond the scope of this book, Chapter 2 describes how a few simple ciphers can be broken using frequency distributions. (For more depth in this area, see [Konh81].)

Protection against statistical analysis can be provided by several means. One way, suggested by Shannon, is by removing some of the redundancy of the lan-