# An Introduction to Cryptanalysis

**Karl A. Siil**

Cryptanalysis is defined as the process of attempting to find a shortcut method, not envisioned by the designer, for decrypting an enciphered message when the key used to encrypt the message is not known. As an example, this paper cryptanalyzes a simple substitution cipher. The methods of cryptanalysis used are the chosen-plaintext attack, the known-plaintext attack, and the ciphertext-only attack, which are defined and discussed in the paper.

## Introduction

Transforming messages into unintelligible data by means of an encryption algorithm is known as *encrypting* or *enciphering*. The algorithm produces an output that cannot be understood, except by someone who knows how to reverse the transformation. Performing such a reverse transformation is known as *decrypting* or *deciphering*.

Encryption has been used by the military to communicate secret information at least as far back as the time of Julius Caesar. Today, networks having huge data bases store information ranging from public-service announcements to highly classified military and government secrets. Encryption is growing in importance on these networks, not only to protect secrets, but also to safeguard such information as a company's intellectual property, or even an individual's medical history.

In the cryptographic community, the original message is commonly called *plaintext* and the output of the encryption algorithm is called *ciphertext*. Because the encryption algorithm is often well known, additional secret information, known as the *key*, is used to perturb the algorithm in some specific manner. Knowledge of the ciphertext and key allows the intended recipient to decrypt the message.

Cryptanalysis is defined as the process of attempting to find a short-cut method—not envisioned by the designer—for decrypting an enciphered message when the key used to encrypt the message is not known.[1]

This is important for governments attempting to maintain national security by analyzing the encrypted transmissions of enemies (and allies). It is also useful as a testing tool. Subjecting a new encryption algorithm to "friendly" scrutiny is far better than using it in the field to code important information, only to discover later that the algorithm is easily "broken" and the information readily disclosed.

Although there are many complex methods currently being used for information-security purposes, the three methods discussed in this paper provide a simple description of the concepts employed in cryptanalyzing encyphered text.

This paper cryptanalyzes a simple substitution cipher. The methods of cryptanalysis used are the *chosen-plaintext attack*, *known-plaintext attack*, and *ciphertext-only attack*. Each of these methods requires less complex data collection by the cryptanalyst than the previous one.

The chosen-plaintext attack allows a choice of which messages the cryptanalyst can encrypt under the unknown secret key. The known-plaintext attack limits the messages to whatever plaintext and corresponding ciphertext the cryptanalyst can gather. Because the plaintext is known, the goal of these two attacks is to find the key.

The ciphertext-only attack requires the least difficult form of data collection. All that must be done is to tap a phone line or intercept a microwave or satellite signal. No plaintext is available to the cryptanalyst. The primary goal of this method is usually to determine the plaintext that produced a given

**Panel 1. Acronyms and Terms**

ASCII—American Standard Code for Information Interchange
character—a letter, numeral, or symbol
cipher—a crypto-system
ciphertext—output of an encryption algorithm
data collection—the gathering of ciphertext and/or plaintext in order to apply a given form of crypt-analysis
decipher—a reversal of the encipher process
decrypt—same as decipher
digram—a set of two characters
encipher—the process of transforming messages into unintelligible data by means of an encryption algorithm
encrypt—same as encipher
frequency analysis—in reference to ciphertext, this process involves counting how many times certain patterns occur and comparing the frequency of occurrences to that of known patterns in English
key—additional secret information used to perturb an encryption algorithm in some specific manner
plaintext—a readable message to be enciphered
substitution cipher—a crypto-system in which each plaintext character is replaced by another character to produce ciphertext
trigram—a set of three characters

ciphertext. A secondary goal—typically more difficult—is to determine the key.

### Algorithm for a Simple Substitution Cipher

The *substitution cipher* is among the simplest of crypto-systems. Though no longer used (and shouldn't be used) to protect data, this cipher can effectively demonstrate the basic principles of cryptanalysis.

As shown in Figure 1, encryption with a substitution cipher involves substituting each plaintext character with another character to produce ciphertext. Usually, the character sets used for plaintext and ciphertext are the same (that is, the English alphabet, ASCII, and so forth), so that the encrypted message can be transmitted using the same medium (paper, data lines) over which the plaintext is transmitted.

Decrypting a message encrypted by a substitution cipher simply reverses the encryption process. Each character of the ciphertext is mapped back to its original value using a substitution table that is the inverse of the table used when encrypting a message.

### Attacking the Cipher

Substitution ciphers are classic crypto-systems having one key that must remain secret if the ciphertext is to be protected. As expected, the secret key in a substitution cipher is the translation table. Attacks against the cipher target this table.

It should be stressed at this point that the cryptanalyst has already assumed being confronted with a given crypto-system—in this case, a trivial one. Such an assumption is not always easy to make. For certain environments, well-known algorithms are defined, and the cryptanalyst can determine which one is being used with relative safety. It can be extremely difficult to know which algorithm is in use, however, if, just by "snooping" around, the cryptanalyst encounters what is believed to be encrypted information.

The methods of cryptanalysis used to find the substitution cipher's translation table are the three attacks mentioned earlier. As previously pointed out, each attack requires less complex data collection by the cryptanalyst than the preceding one. Intuitively, the more data that is required for the attack, the more difficult the attack is to launch.

The chosen-plaintext attack requires the most data collection effort, as the cryptanalyst might require access to the actual encryption equipment. An alternate collection method is to generate a message and inject it into the encryption system so that the message is encrypted, transmitted, and subsequently intercepted. This message could take the form of several fictional bank transactions. Specific amounts could be deposited into, and withdrawn from, certain accounts. Starting with the assumption that the message is part of the ciphertext, the cryptanalyst follows the steps of the chosen-plaintext attack. By substituting specific transaction values, the attacker can attempt to exploit any weaknesses of the encryption algorithm.

Use of the known-plaintext attack precludes choosing the plaintext. The cryptanalyst is left with whatever information can be obtained. This attack implies that the attacker has no way to insert input into the actu-

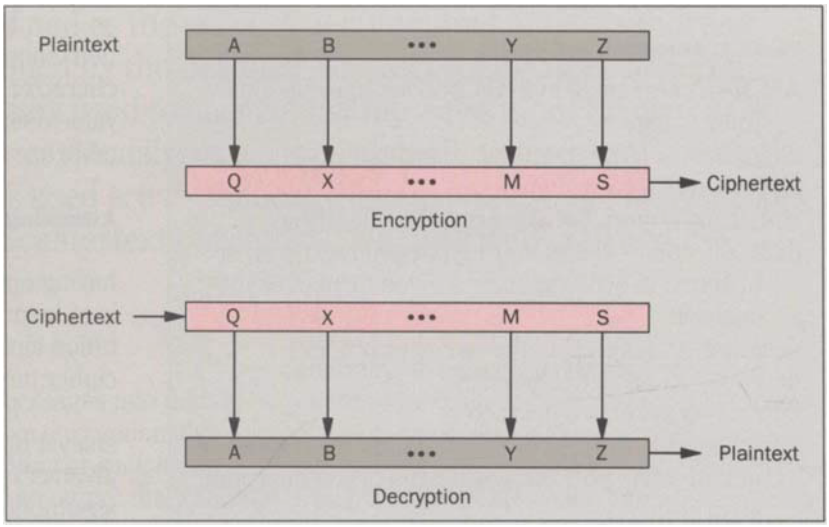


**Figure 1. Encryption with a substitution cipher involves substituting each plaintext character with another character to produce ciphertext. Usually, the character sets used for plaintext and ciphertext are the same, so that the encrypted message can be transmitted using the same medium over which the plaintext is transmitted. Decrypting a message encrypted by a substitution cipher simply reverses the encryption process. Each character of the ciphertext is mapped back to its original value using a substitution table that is the inverse of the table used when encrypting a message.**

al encryption equipment. In the case of the fictional bank transactions, the cryptanalyst has either no account at the bank or insufficient knowledge about the bank's electronic transaction formatting. However, the cryptanalyst still has a body of plaintext to work with. The reason for this could be that the same encrypted channel (tapped by the cryptanalyst) is used for the bank's direct mail, and the attacker gets that mail all the time. The attacker assumes, for instance, that the bank statement is somewhere in the data. The attacker knows what that is and, therefore, has some "known plaintext."

The ciphertext-only attack is the most restrictive because no plaintext is available. This attack, however, is easiest to launch. The only requirement is to tap a phone line or intercept a microwave or satellite signal. Continuing with the bank-transaction scenario, the cryptanalyst may have decided to attempt to harm the bank (for whatever reason), and has discovered one of the channels that the bank uses for encrypted transmission.

**The Chosen-Plaintext Attack.** Because the cryptanalyst is given the most initial information, this approach promises to be the most productive. Also, this method of attack is not altogether unrealistic. In all probability, a determined attacker can either gain access to the crypto-system or generate a message that, it is assumed, resides in the ciphertext that is being analyzed.

For this attack, any message the attacker considers necessary may be enciphered with some unknown key. The attacker then compares the plaintext with the ciphertext and searches for patterns that may lead to determining the mechanics of the algorithm. Given this process, the obvious choice when attacking a substitution cipher is to input whatever alphabet the algorithm uses as the initial plaintext. In the case of English, we would encrypt "ABCD...XYZ," which would result in the output of ciphertext that *is* the translation table. This is shown below:

```
(input)  ABCDEFGHIJKLMNOPQRSTUVWXYZ
(output) ZYXWVUTSRQPONMLKJIHGFEDCBA
```

This substitution cipher is now "cracked." We no longer need to see the plaintext for any given ciphertext. Given the above translation table, which happens to be the English alphabet in reverse, we can decrypt any ciphertext we receive that has been encrypted by this crypto-system.

Although there are more than $4 \times 10^{26}$ different substitution ciphers using the English alphabet, the vulnerability of this type of crypto-system to the chosen-plaintext attack makes it a poor choice to protect information. With as few as 25 characters of chosen input, the protection of the cipher is entirely lost.

**The Known-Plaintext Attack.** This approach is more restrictive then the chosen-plaintext method. The cryptanalyst is not allowed to encipher just any message. Instead, a set of plaintext messages, and their corresponding ciphertexts, are made available to the cryptanalyst.

The increased effort level is partially due to no guarantee that a sufficient amount of plaintext is avail-

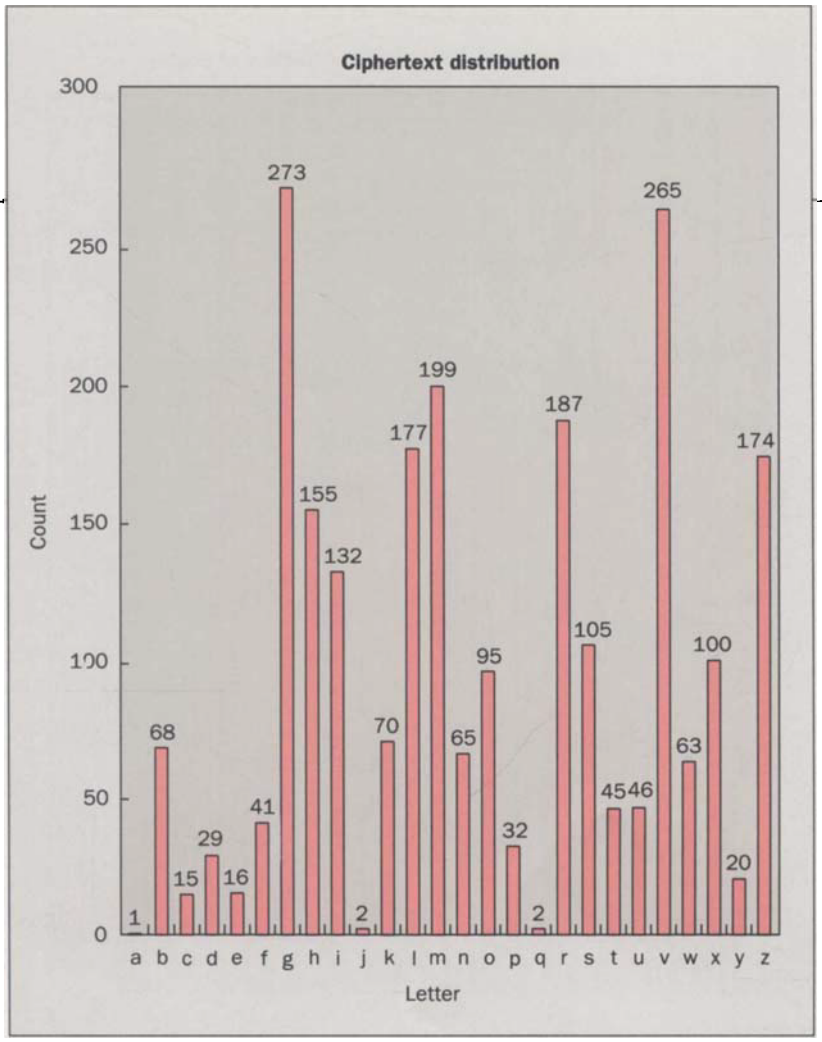


**Figure 2. The ciphertext-only attack on a substitution cipher is *frequency analysis* of the ciphertext. This involves counting how many times certain patterns occur and comparing the frequency of those occurrences to the frequency of known patterns in English. The frequency distribution of roughly 3,000 characters of ciphertext is shown.**

able to recover the entire translation table or key. Another cause of the reduced efficiency of this attack is that the cryptanalyst, having no control over the available plaintext, may receive redundant messages, providing little additional insight about the key.

Given the following short plaintext message and its associated ciphertext, the mechanics of a substitution-cipher attack are quite simple:

*Plaintext*: Transforming messages into unintelligible data via an encryption algorithm is known as encrypting or enciphering. The algorithm produces an output that cannot be understood, except by someone who knows how to reverse the transformation. Performing this reverse transformation is known as decrypting or deciphering.

*Ciphertext*: Gizmhulinrmt nvhhztvh rmgl fmrmgvoortryov wzgz erz zm vmxibkgrlm zotlirgsn rh pmldm zh vmxibkgrmt li vmxrksvirmt. Gsv zotlirgsn kilwfxvh zm lfgkfg gszg xzmmlg yv fmwvihgllw, vcxvkg yb hlnvlmv dsl pmldh sld gl ivevihv gsv gizmhulinzgrlm. Kviulinrmt gsrh ivevihv gizmhulinzgrlm rh pmldm zh wvxibkgrmt li wvxrksvirmt.

The cryptanalyst constructs the translation table by determining which characters have been translated to other characters. For example, T becomes G, R becomes I, and A becomes Z. Using the preceding cipher text, it can be determined what all but three ciphertext characters (A, J, and Q) translate to in plaintext. The resulting translation table is composed as follows:

```
(input)   ABCDEFGHIJKLMNOPQRSTUVWXYZ
(output)  _YXWVUTSR_PONMLK_IHGFEDCBA
```

It is highly likely that any ciphertext encrypted using this crypto-system can be decrypted, because the remaining *mystery letters* (J, Q, and Z) are not widely used in English. In actuality, the remainder of the translation table will probably be determined by decrypting future messages.

Once again, the vulnerability of this type of crypto-system to attacks that include plaintext makes it a poor choice for protecting sensitive data. In the case of the known-plaintext attack, only a few hundred characters are needed to all but totally crack the cipher. Though the translation table is not fully resolved by this attack, what little text remains secret is of no consequence to the cryptanalyst.

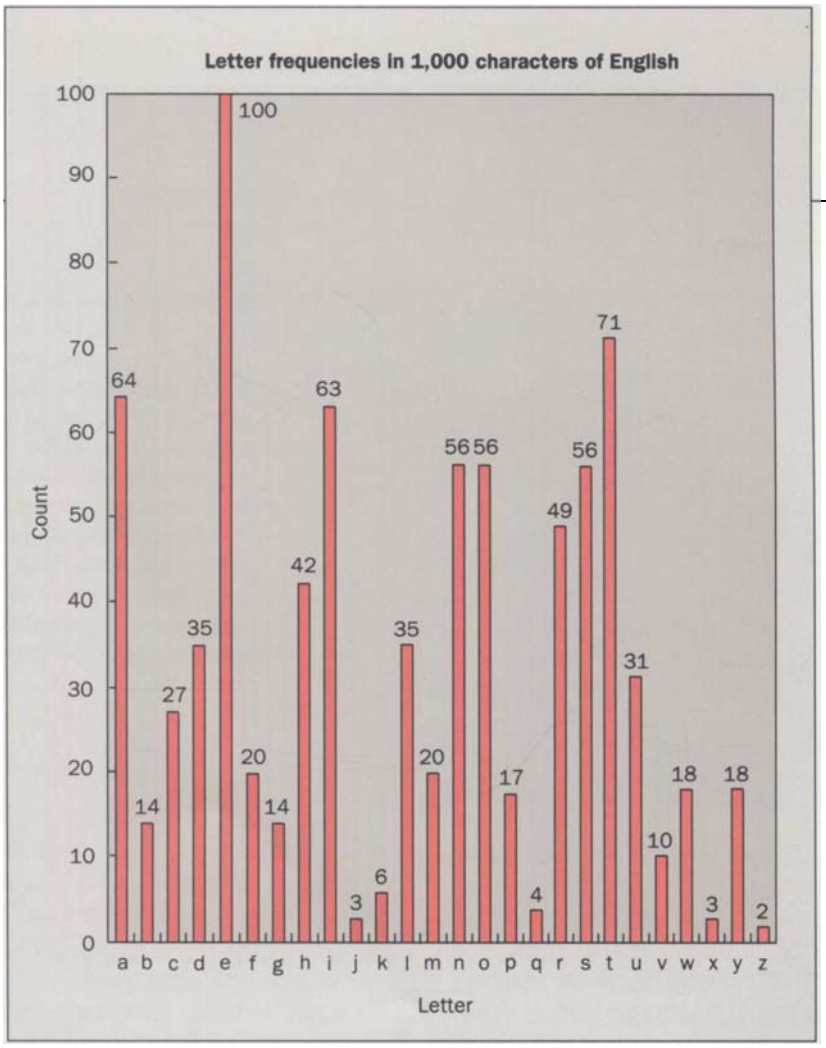


**Figure 3. As explained by the caption for Figure 2, the ciphertext-only attack on a substitution cipher is *frequency analysis* of the ciphertext. The frequency distribution of roughly 3,000 characters of ciphertext is shown in Figure 2. In contrast, the frequency distribution of characters in English is illustrated by this bar graph.**

**The Ciphertext-Only Attack.** For this approach, the only information given to the cryptanalyst is ciphertext generated by the crypto-system. It is further assumed that the cryptanalyst has determined (or decided to assume) that the crypto-system in use is a substitition cipher.

The classic ciphertext-only attack on a substitution cipher is *frequency analysis* of the ciphertext. This process involves counting how many times certain patterns occur and comparing the frequency of those occurrences to the frequency of known patterns in English. The patterns analyzed are usually *characters*, *digrams*, and *trigrams*. The latter two are simply pairs and triples of characters, respectively. Only characters, however, are analyzed and discussed in this paper.

Figure 2 shows the frequency distribution of roughly 3,000 characters of ciphertext given to the cryptanalyst. In contrast, the frequency distribution of characters in English is shown in Figure 3.[2]

The analysis can be continued by assuming that G or V in the ciphertext translates to E in plaintext. This is because E is the most popular letter in English, and the other two letters form a significant fraction of the ciphertext. It is also likely that whichever letter does not translate to E translates to T, because T is the second most popular English-alphabet letter.

By substituting our assumptions into the ciphertext, and then making calculated guesses about the other letters from the resulting partial translations, the translation table slowly evolves. When the table is complete, it is found that the resulting plaintext (the "Introduction" section of this paper) has roughly the same letter distribution, overall, as does the English language. This phenomenon is illustrated by Figure 4.

## Conclusion

This paper introduces the science of cryptanalysis and provides an example, by cryptanalyzing a simple substitution cipher, of rudimentary use of its techniques. The methods of cryptanalysis used are the chosen-plaintext attack, known-plaintext attack, and ciphertext-only attack.

Under the chosen-plaintext and known-plaintext attacks, the cipher was easily determined by cryptanalysis. Using input plaintext, consisting of the English alphabet in the chosen-plaintext attack, the key was recoverable with only 26 characters of ciphertext. Under the

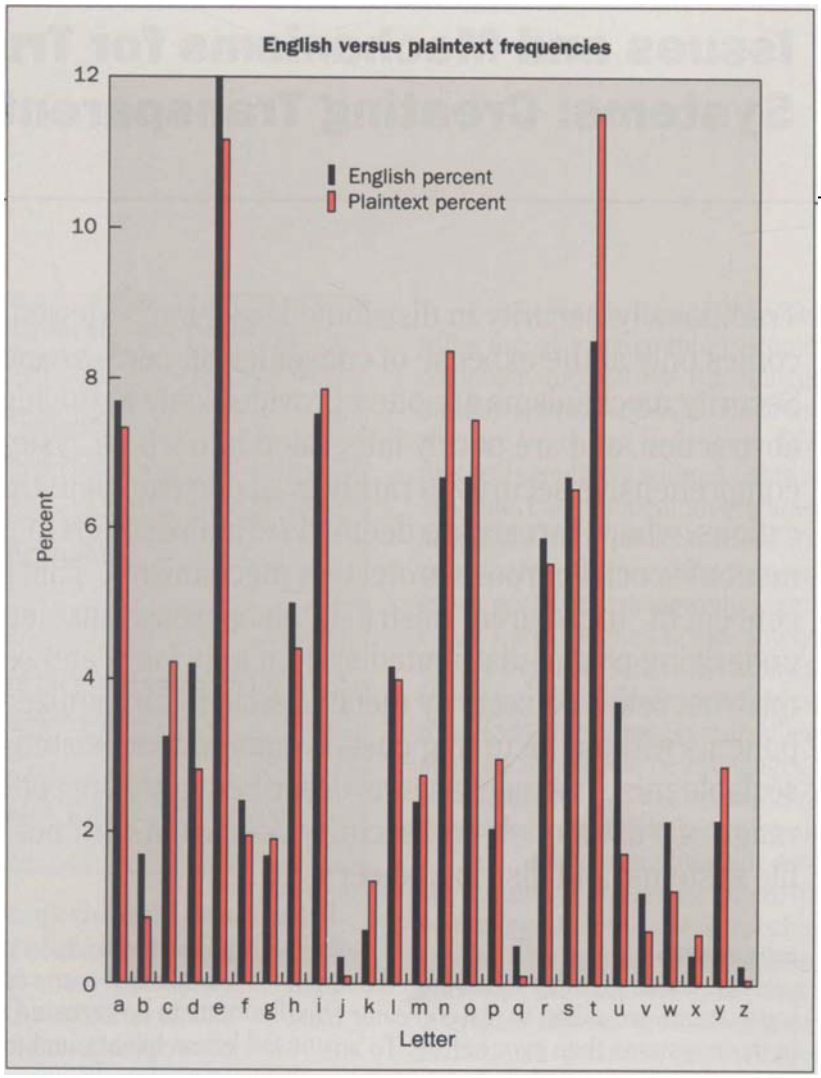


**Figure 4. By substituting certain assumptions into ciphertext, and then making calculated guesses about other letters from the resulting partial translations, a translation table evolves. When the table is complete, the resulting plaintext (the "Introduction" section of this paper) has roughly the same letter distribution, overall, as does the English language. This phenomenon is illustrated by the bar graph.**

known-plaintext attack, all but three elements of the translation table can be ascertained from only a few hundred characters of plaintext, along with the plaintext's associated ciphertext.

Under the ciphertext-only attack, a frequency analysis of the data yielded a similarity with the frequency distribution in English. Using this information, a series of increasingly educated guesses, with as few as 3,000 input characters, eventually leads to the translation table.

As stressed in this paper, substitution ciphers are antiquated, easily cracked crypto-systems. The example of such a system in this discussion is solely for demonstration purposes.

**References**

1. G. J. Simmons (ed.), *Contemporary Cryptology: The Science of Information Integrity*, IEEE Press, Piscataway, New Jersey, 1992.
2. D. Welsh, *Codes and Cryptography*, Oxford University Press, New York City, 1988.

*(Manuscript approved August 1994)*

***Karl A. Siil*** *is a member of the technical staff in the PersonaLink Services department at AT&T EasyLink Services in Lincroft, New Jersey. He is responsible for system-engineering globalization of PersonaLink services and the use of international cryptographic standards. He has also worked on authentication subsystems, including cryptographic security regimes. Mr. Siil has two degrees in electrical engineering, a B.E.E. from The Cooper Union and an M.E.E. from Manhattan College, both in New York City. He joined AT&T in 1987.*