A homophonic cipher for computational cryptography*

by FRED A. STAHL

University of Illinois at Urbana-Champaign
Urbana, Illinois

INTRODUCTION

Computational cryptography deals with the storage and processing of sensitive information in computer systems by enciphering. Sensitive information is information that for one reason or another must be protected from being disclosed to individuals without proper authorization. The need for systems to be secure from unauthorized access to sensitive information has been well documented.[1-7] Cyrptographic techniques appear to be one of the most simple and secure methods of providing this much needed protection.

In the next section the requirements for computational cryptography will be discussed, especially with regard to the differences from communication cryptography; second, a short review of the cryptographic techniques that have been suggested for computational cryptography; third, a review of some standard cryptanalysis techniques; fourth, a homophonic cipher will be described; and finally, some additional problems associated with computational cryptography will be discussed.

REQUIREMENTS FOR COMPUTATIONAL CRYPTOGRAPHY

Cryptography has been used for ages as a technique for concealing the informational contents of messages. The use of cryptanalytic techniques combined with the availability of high speed electronic equipment have made both encipherment and cryptanalysis very sophisticated endeavors indeed. For this reason, simple ciphers can no longer be used to provide a great deal of security. This carries over to computational cryptography also. Moreover, there are a number of desirable features that any ciphers to be used in computer systems should have. These include:

1. Encipherment and decipherment should be simple. That is, the computational complexity of enciphering and deciphering information should be minimal.
2. The effective ability to not be broken should be high. As the sensitivity of the information increases a cipher that is more difficult to break should be available for use.
3. The cipher should be independent of message length. Messages can be even one character in length.
4. Small errors should not cause large losses of information. If a small amount is ciphered incorrectly it should not make a large block of information undecipherable.
5. Information should remain integral through editing of the ciphered text. The cipher should be such that the normal editing functions of deleting, inserting and moving strings of information (such as would occur in a dynamic data-base) can be performed without destroying the information contained in the ciphered text.
6. The cipher should not increase the length of the message excessively.
7. The cipher should be of such a nature that security can be maintained even though the cryptanalyst knows the ciphering system but not the key.

There are in addition to these desirable features a number of problems peculiar to computational cryptography. These will be discussed in the last section.

CRYPTOGRAPHIC TECHNIQUES THAT HAVE BEEN SUGGESTED FOR COMPUTATIONAL CRYPTOGRAPHY

The major thrust, so far, in computational cryptography has been to provide a cipher key to the computer and let it encipher information that enters the system, and also decipher and reencipher information as it is processed by the machine. The ciphers suggested include: simple substitution schemes, arithmetic schemes (i.e., adding or multiplying by a constant, or base conversion), logical schemes (e.g., exclusive-or), and transposition schemes. These have been described quite thoroughly by Krishnamurthy,[8] and Van Tassel.[9,10]

Some of these ciphers are too easily broken by the most elementary cryptanalytic techniques and others are too computationally complex to be implemented for use in computers. All fail to have some of the desirable features listed above.

* The work presented here was performed at the Coordinated Science Laboratory of the University of Illinois at Urbana-Champaign. It was supported in part by the Joint Services Electronics Program (U.S. Army, U.S. Navy, U.S. Air Force) under Contract No. DAAB-07-72-C-0259.

STANDARD CRYPTANALYTIC TECHNIQUES

It is not the purpose here to describe in great detail the cryptanalytic techniques that can be used to break a cipher. The reader interested in this topic is referred to Gaines.[11] It is only intended to suggest what types of information remain visible after the original text has been enciphered.

For the most part cryptanalytic clues are gained from the information inherent in the structure of the language such as frequency information which is extremely hard to remove. Simple substitution, for example, leaves for easy analysis all single letter, and multiple letter frequencies, doublet and reversal frequencies, as well as contact variety information. Figure 1 is representative of this information for a standard text.

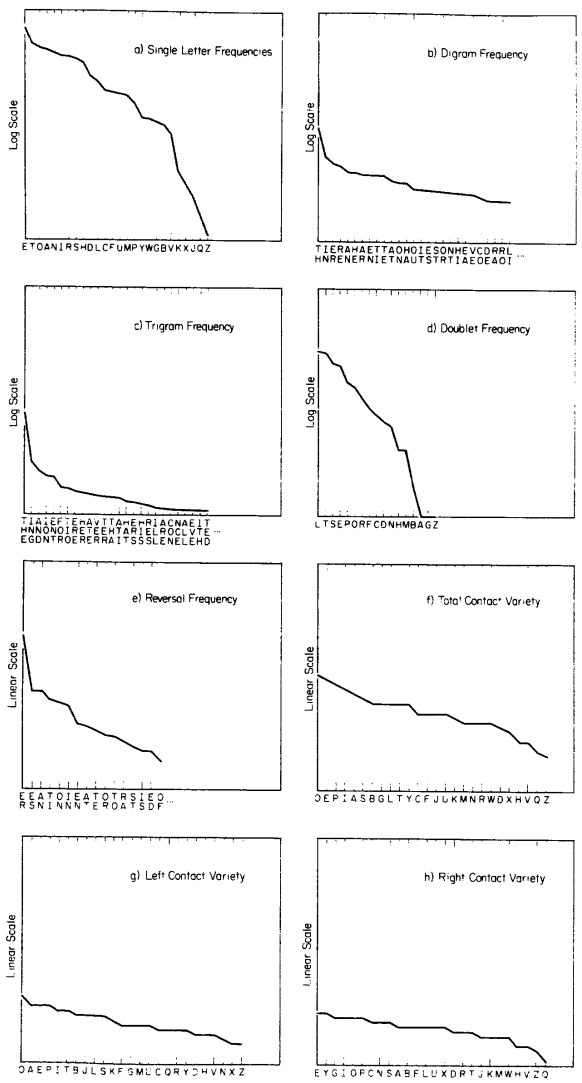

Figure 1—Various frequency distributions for English

HOMOPHONIC CIPHERS

A homophonic cipher is a substitution cipher in which a given character may have any of a number of different representations. Figure 2 gives one such cipher and a sample message using it. Note that the cipher-text for E, for instance, varies from substitution to substitution. Kahn[12] notes that the first known Western instance occurs in a cipher that the Duchy of Mantua prepared in 1401 for correspondence with one Simone de Crema. Each of the plaintext vowels has several possible equivalents.... "That the homophones were applied to vowels, and not just indiscriminately, indicates a knowledge of at least the outlines of frequency analysis."

Obviously, the more ciphertext symbols relative to plaintext symbols the easier it is to disguise the structural properties of the plaintext through enciphering. Each plaintext symbol could have many ciphertext encipherings. In order to illustrate this consider a plaintext alphabet of 26 symbols and a ciphertext alphabet of 1024 symbols (10 bits). Initially each plaintext symbol would have as many ciphers directly proportional to its frequency in the language (see Figure 3). Notice that the single letter frequency of the ciphertext is nearly constant (compare Figure 4 to Figure 1a) relative to its frequency in the plaintext thereby not providing any single letter frequency information for the cryptanalyst.

However, as noted above, the cryptanalyst uses other techniques for breaking codes. The cryptanalyst looks for the most deviant structural features first. Consider, for example, the abnormal reversal frequency of ER in Figure 1e. If this reversal can be located in the ciphertext the cryptanalyst is much closer to breaking the code. In contrast, he would not look for the reversal OF since there are many other reversals with nearly the same frequency occurrence. The crucial point to remember is that it is only the abnormal or deviant frequencies that can be used for clues. Clearly, if all measurements of the ciphertext yielded nearly flat distributions as in Figure 4 there would be no information gained from those measurements.

Key

Plaintext:	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
Ciphertext:	C R Y P T O G R A M 5 6 7 8 9 B D E F H I J K L N Q
	1 2 3 4 S
	U V W X
	Z

Message

Plaintext:	THIS IS A SECRET MESSAGE
Ciphertext:	HRAF AF C FTYE2H 7VFF1GZ

FP-3445

Figure 2—A simple homophonic substitution cipher

Plaintext Symbol:	A	B	C	D	E	F	G	H	I	J	K	L	M
Number of Cipher Symbols:	81	13	31	43	133	29	14	61	71	2	5	38	27

	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
	15	85	22	2	70	65	93	28	10	15	3	15	1

FP-3448

Figure 3—A homophonic substitution cipher generated in direct proportion to frequency in plaintext

Now let us generalize the homophonic technique to flatten the other curves illustrated in Figure 1. This is easier said than accomplished, since adjusting one curve to be flatter will generally result in making another one more curved.

Let us return to the earlier generalized homophonic enciphering hypothesis and modify it somewhat to allow for more flexibility. Before we wanted the curve resulting from single letter enciphering to be nearly flat. If we lessen this restriction somewhat and allow for some distribution but not nearly as much we can avoid the abnormal or deviant frequencies that are normally used for clues on all the measurements found in Figure 1. Figure 5 gives one such homophonic cipher.

The technique used to generate this code is fairly simple. The frequencies are adjusted as for the code in Figure 3, and the corresponding frequency charts can be generated for the other measurement of the ciphertext. Next, the most deviant frequency in any measurement is examined. If, for example, it is a reversal frequency the corresponding number of cipher symbols used for each of the constituent symbols is raised accordingly. The process is repeated until the frequency curves are satisfactory. If there is no convergence the process can be started over again taking care to choose different measurements first. If after a number of attempts are made appropriate curves cannot be gotten, it might be necessary to increase the number of ciphertext symbols by increasing the number of bits used to represent each symbol.

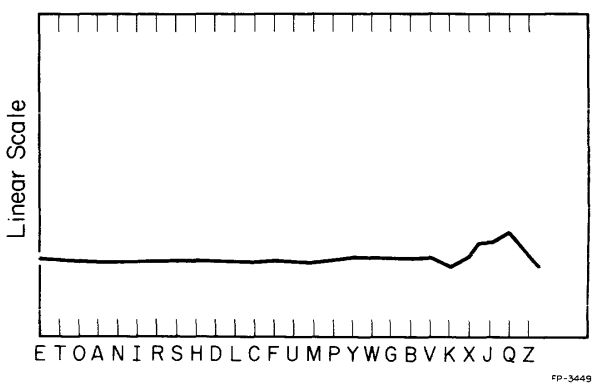

Figure 4—Single cipher frequency for homophonic cipher

Plaintext Symbol:	A	B	C	D	E	F	G	H	I	J	K	L	M
Number of Cipher Symbols:	73	10	24	34	129	23	10	52	77	2	4	43	22

	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
	27	89	19	12	80	72	87	45	8	13	2	13	2

FP-3450

Figure 5—A generalized homophonic cipher

A cipher successfully generated in the manner described has all the desirable features for computational cryptography set forth above. An extremely simple enciphering and deciphering scheme can be used since this is a substitution type cipher. A sample scheme will be given. With regard to the ability to adjust the cipher to meet security needs only increase the number of bits used to represent each ciphertext symbol (effectively, increasing the ciphertext alphabet). Since there is one ciphertext symbol for each plaintext symbol the messages can be of arbitrary length. An error in one symbol does not extend errors even to adjacent symbols of the message; thereby keeping losses of information to a minimum. Note in particular that since it is a substitution cipher the 'integrity through editing' condition is met. That is, strings (including individual characters) of enciphered message may be moved with respect to each other without going through a deciphering and reenciphering process. This property makes it invaluable for large dynamic databases.

The length of the message only increases with the security needed. For a typical low security cipher 8-bits should be sufficient for a 64 symbol plaintext alphabet. A homophonic cipher can effectively destroy all standard language frequency information as shown in Figure 1. In addition, information in ciphered form may be received by the computer from a terminal and be edited without it ever being deciphered at the central facility. As mentioned earlier the device to encipher the plaintext message need not be very complex. Consider a key of 256 characters; each of the 64 characters appears in the key the number of times desired for the particular application (see Figure 6). The key is loaded into a 256 word memory. Deciphering consists of returning the contents of the address specified by the 8-bit cipher. Enciphering involves generating an address randomly and then searching sequentially until a matching character is found and then transmitting its address (see Figure 7).

The amount of secrecy needed can be controlled by the number of bits. So, for instance with 9 bits there would be 418 remaining bit patterns; with 10 bits it would be 984, etc. Each additional bit increases the security.

ADDITIONAL PROBLEMS ASSOCIATED WITH COMPUTATIONAL CRYPTOGRAPHY

There are a number of additional problems associated with keeping sensitive information secure in computer systems. These include the following:

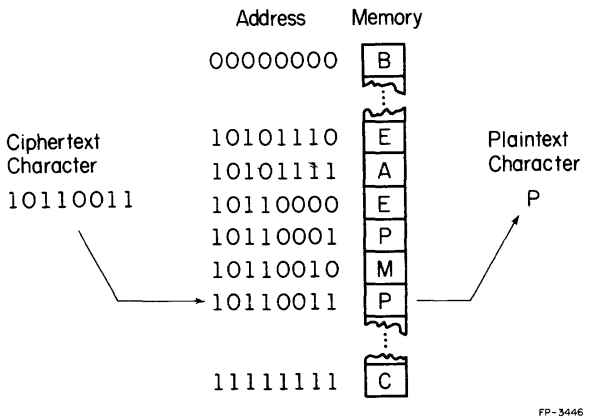

Figure 6—Deciphering the homophonic cipher

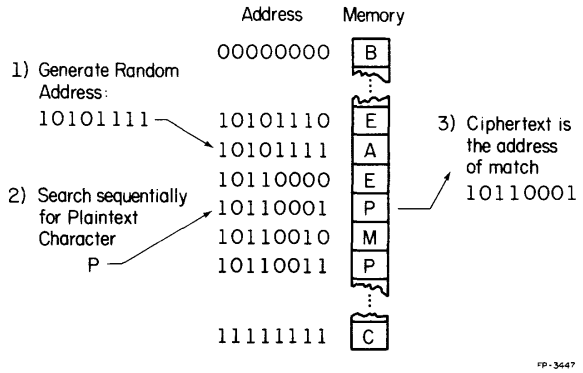

Figure 7—Enciphering the homophonic cipher

1. *The decoded message problem.* This comes about when a block of decoded or unenciphered message is known by the analyst. With this type of information available very few cipher systems are safe.
2. *The limited syntax problem.* When dealing with limited languages such as programming languages the analyst can break the cipher by knowing the restrictive properties of the language involved.
3. *The arithmetic problem.* If the ciphered text is not to be decoded inside the computer arithmetic operations cannot be performed.
4. *The overlapping access problem.* When different individuals have access to the same ciphered data and yet do not want common access to other enciphered data.

It is hoped that this paper will influence in some way the designers of future computer systems by showing that simple techniques can be used for effectively limiting the access of information to only those who should have access to it. A justification of the effectiveness of this homophonic cipher system on a mathematical basis using techniques developed by Shannon[13] will be described in a subsequent paper.

ACKNOWLEDGMENTS

The author wishes to thank Chung Laung Liu and James Studier for their suggestions and criticisms of the manuscript during its preparation and Greg Michael for his program to compute the various frequencies for the figures that appear.

REFERENCES

1. Ware, W. H., "Security and privacy in computer systems" *Proceedings Spring Joint Computer Conference*, pp. 279-282, 1967.
2. Ware, W. H., "Security and Privacy - Similarities and Differences", *Proceedings Spring Joint Computer Conference*, pp. 287-290, 1967.
3. Peters, B., "Security Considerations in a Multi-Programmed Computer System", *Proceedings Spring Joint Computer Conference*, pp. 283-286, 1967.
4. Petersen, H. E., Turn, R., "Systems Implications of Information Privacy", *Proceedings Spring Joint Computer Conference*, pp. 291-300, 1967.
5. *The Computer and Invasion of Privacy*, hearings before a Subcommittee on Government Operations, House of Representatives, July 26-28, 1966, Government Printing Office.
6. Harrison, A., *The Problem of Privacy in the Computer Age - An Annotated Bibliography*, Rand Corporation Memorandum RM=5495-PR/RC, December 1967.
7. Harrison, A., *The Problem of Privacy in the Computer Age - An Annotated Bibliography Vol. 2*, Rand Corporation Memorandum RM=5491/1-PR/RC, December 1969.
8. Krishnamurthy, E. V., "Computer Cryptographic Techniques for Processing and Storage of Confidential Information", *International Journal of Control*, Vol. 12, No. 5, pp. 753-761.
9. Van Tassel, D. L., "Cryptographic Techniques for Computers", *Proceedings Spring Joint Computer Conference*, pp. 367-372, 1969.
10. Van Tassel, D. L., "Advanced Cryptographic Techniques for Computers", *Communications of the ACM*, Vol. 12, No. 12, December 1969, pp. 664-665.
11. Gaines, H. F., *Cryptanalysis*, Dover, 1956.
12. Kahn, D., *The Codebreakers*, Macmillan, 1967.
13. Shannon, C. E., "Communication Theory of Secrecy Systems", *Bell System Technical Journal*, Vol. 28, No. 4, October 1949, pp. 656-715.