EQUIVOCATIONS FOR HOMOPHONIC CIPHERS

Andrea Sgarro
Istituto di Matematica
Università di Trieste
34100 Trieste (Italy)

Abstract. Substitution ciphers can be quite weak when the probability distribution of the message letters is distinctly non-uniform. A time-honoured solution to remove this weakness is to "split" each high-probability letter into a number of "homophones" and use a substitution cipher for the resulting extended alphabet. Here the performance of a homophonic cipher is studied from a Shannon-theoretic point of view. The key and message equivocations (conditional entropies given the intercepted cryptogram) are computed both for finite-length messages and "very long" messages. The results obtained are strictly related to those found by Blom and Dunham for substitution ciphers. The key space of a homophonic cipher is specified carefully, so as to avoid misunderstandings which appear to have occurred on this subject.

Work done within the research program of GNIM-CNR.

T. Beth, N. Cot, and I. Ingemarsson (Eds.): Advances in Cryptology - EUROCRYPT '84, LNCS 209, pp. 51-61, 1985.
© Springer-Verlag Berlin Heidelberg 1985

1. Introduction.

Simple substitution ciphers (s.s.c.'s) are probably the oldest type of ciphers put to work, and yet they are still in good health in the form of (individually weak) components of (hopefully good) complex cipher system (e.g. the Data Encryption Standard). The key of a s.s.c. is a permutation of the message-letter alphabet $A=\{a_1,a_2,\dots,a_s\}$, $s\geq 2$; once a key is chosen each single letter output by the message source is replaced by its substitute. S.s.c.'s have been studied rather deeply in the last decade; cf. /1, 2, 3/; in the first two papers the strength of a cipher is assessed by evaluating the equivocations ("uncertainties") on side of the spy who has intercepted a cryptogram (key equivocation or message equivocation, according whether the spy is interested in finding out the correct key or the correct message); in /3/ the error probabilities are evaluated when the spy uses the best statistical procedure to recover the correct key or message from the intercepted cryptogram. Further work on s.s.c.'s is done in /4/, which contains a discussion on the role of the "Shannon-theoretic approach" to cryptography and, more generally, on the relevance of purely statistical cryptographic models.

A s.s.c. is very weak when the probability distribution (p.d.) P ruling the message source, which we assume to be memoryless and stationary, is distinctly non-uniform; ($P=\{p_1,p_2,\dots,p_s\}$, $p_i>0$, $\sum p_i=1$; unspecified summations are meant over all values of the index). A time-honoured solution to remove this weakness is to make use of a cryptogram-letter alphabet C of size t larger than s, the size of the message-letter alphabet; for example, the letters of C might be the ordered couples of message letters. Then any large probability p_i can be broken down by associating to the corresponding letter a_i many possible cryptogram substitutes, t_i, say: each time a_i occurs in the message one of these is chosen at random and actually substituted for a_i. The resulting cipher is called a homophonic cipher (or, rather, a simple, that is single-letter, homophonic cipher; a more formal description is given below). Homophonic ciphers, which are a generalization of s.s.c.'s (refound for $t_i=1$, $1\leq i\leq s$, that is, essentially, when $A=C$) have been recently

studied in /5/. In this paper we take the equivocation approach to assess the strength of homophonic ciphers, thereby generalizing the work done in /1/ and /2/ for s.s.c.'s.

A mathematical tool which we shall use is the notion of an exact type. Consider A^n, the set of the s^n sequences of length n built over A; an exact type (of order n over A) is a subset of A^n made up of a sequence together with all its permutations. Of course, if $(n_1,n_2,\ldots,n_s)$ is the composition of any of these sequences, n_i being the number of occurrences of letter a_i, $(n_i \geq 0, \sum n_i = n)$, the size of the corresponding exact type is the multinomial coefficient $n!/n_1!n_2!\ldots n_s!$. Statisticians will recognize here an obvious link with the notion of sufficient statistic; we simply stress that the sequences of an exact type have all the same probability. A powerful technique based on asymptotically tight bounds for the size and the probability of exact types has been made popular in the circle of information theorists by the fundamental textbook /6/. This technique is applied in /3/ to the error probability approach to s.s.c.'s and in/7/ to the equivocation approach to the same ciphers.

Before going to mathematical developments, we have to give a more formal description of a homophonic cipher. Two alphabets, A and $C = \{c_1,c_2,\ldots,c_t\}$, $t \geq s$, are given, C being the cryptogram-letter alphabet; also s integers are given which sum to t: $t_1,t_2,\ldots,t_s$, $t_i \geq 1$, $\sum t_i = t$. A key is specified by giving s disjoint subsets of C of size $t_1,t_2,\ldots$ and t_s, respectively. Each time letter a_i is output by the message source, one of the t_i letters of the i-th subset is chosen with (conditional) probability $1/t_i$ and is substituted for letter a_i. The knowledge of the key is enough to reconstruct the correct message from any of the possible corresponding cryptograms. Before transmission begins, a key is chosen at random and independently of the message output by the source; the key is communicated to the legitimate receiver via a secure special channel; ("at random" means that the key is a uniform random variable, or r.v., over the set of all possible keys). The cryptogram is derived from the message and sent over the normal unsafe channel, where it is intercepted by the spy.

We find it convenient to give a more careful description of the key of a homophonic cipher. Such a key can be represented by a sequence in

A^t with composition $(t_1,t_2,\ldots,t_s)$ (alphabet letter a_i occurs t_i times, $\sum t_i=t$). The meaning of this representation is that if a_i is the j-th component of the sequence, then c_j is a possible substitute for a_i under the given key. We shall actually identify each possible key with the corresponding sequence in A^t, so that for us the set of keys will be an exact type in A^t. Clearly the number of all keys for a homophonic cipher $(A, C; t_1,t_2,\ldots,t_s)$ is the multinomial coefficient $t!/t_1!t_2!\ldots t_s!$

As shown in /5/, a homophonic cipher induces a s.s.c. in a quite natural way. A presentation follows which suits our purposes. Take an extended alphabet $U=\{u_1,u_2,\ldots,u_t\}$ with the same size t as the cryptogram alphabet C; although it would not be restrictive to take $U=C$ we keep them separate for the sake of notational clarity. The elements of U will be denoted at places by symbols like a_{ij}, $1\leq i\leq s$, $1\leq j\leq t_i$; in other words in U each message letter a_i is duplicated t_i times: the letters a_{ij} are called the homophones of letter a_i. No ambiguity should result from the fact that the letters of U have two names, e.g. u_1 is also called a_{ij} for some i and some j. A dummy memoryless and stationary source with alphabet U, called the extended source, is now built in the following way: each time the message source outputs a letter a_i, $1\leq i\leq s$, the extended source outputs a letter a_{ij}, $1\leq j\leq t_i$, with (conditional) uniform probability; then the (absolute) probability of letter a_{ij} is p_i/t_i. We call P* the p.d. made up of these probabilities; P* rules the statistical behaviour of the extended source. Note also that the output of the message source is a deterministic function of the simultaneous output of the extended source.

Take now the s.s.c. (U, C), whose t! keys can be represented (cf. above) as sequences in U^t where each letter occurs exactly once. To any key for (U, C) we can associate a key for $(A, C; t_1,t_2,\ldots,t_s)$ replacing each a_{ij} in the U^t sequence by a_i. A homophonic cipher can be put to work in the following way, which is readily shown to be equivalent to the original description. A key is chosen for the s.s.c. (U, C). The message source is set going together with the extended source syncronized with it. The key is applied to the extended message to give the cryptogram. From this key a "short" key can be obtained as above to be communicated

to the legitimate receiver. The "short" key applied to the cryptogram does not allow the legitimate receiver to recover the extended message, but it does allow him to recover the original message over A, which is what he needs to know. The key for the s.s.c. (U, C) is a uniform r.v. K* with t! values, while the "short" key for the homophonic cipher is a uniform r.v. K with $t!/t_1!t_2!...t_s!$ values. We shall also write $K^*=(K, J)$, where J is again a uniform r.v., this time with $t_1!t_2!...t_s!$ values, which identifies K* once K is known; note that K and J are independent, as it appears from the values of the respective probabilities. K, J and K* will be referred to as the actual key, the supplementary key and the extended key. We stress the distinction between K and K*: it is the former which is the "true" key of the homophonic cipher, while K* contains the "redundant information" J; cf. the discussion at the end of section 4.

Let the r.v.'s M_n, U_n and C_n denote the first n letters output by the message source, the extended source and the cryptogram source, respectively. Some relations for relevant entropies are already implicit from the foregoing: $H(M_n|U_n)=0$, $H(M_n|K,C_n)=0$, $H(M_n|K)=H(M_n|K^*)=H(M_n)$, $H(K,J)=H(K)+H(J)$, etc. In the following section we shall assess the performance of a homophonic cipher by evaluating its equivocations.

2. The equivocations.

The equivocations of interest are: $H(K|C_n)$, the key equivocation, $H(K|M_n,C_n)$, the key appearance equivocation, interesting in the case of "chosen plain-text attacks", and, most important of the three, $H(M_n|C_n)$, the message equivocation. Since (U, C) is a s.s.c. we already know a lot about its own equivocations, $H(K^*|C_n)$, $H(K^*|U_n,C_n)$ and $H(U_n|C_n)$; cf. /1,2,7/. Only the first will be needed. Its value is

$$(1) \qquad H(K^*|C_n) = \log A + \sum_{\underline{r}} P^{*n}(T_{\underline{r}}) \log \frac{\sum_{\underline{Q}} Q^n(\underline{u}_{\underline{r}})}{P^{*n}(\underline{u}_{\underline{r}})}$$

where $A=A(P,t_1,t_2,...,t_s)=d_1!d_2!...d_h!$, h being the number of distinct probabilities appearing in P*, the first d_1 times, the second d_2 times,

etc.; $d_1+d_2+\ldots+d_h=t$; the r-summation is extended over all exact types T_r in U^n; $\underline{u}_r$ is any sequence in T_r; the Q-summation is extended over all p.d.'s Q which are obtained by permuting the components of P*, including P* itself: these p.d.'s are only t!/A owing to ties in the components of P*; of course Q^n is the memoryless extension of Q over U^n. The term log A is a constant; it is certainly non-zero for a strictly homophonic cipher (one for which $t\geq s+1$). The second term goes to zero and it is exactly zero when P* is uniform and A achieves its maximum value log t! (cf. also section 3).

Some simple identities are helpful. For example: $H(K^*|C_n)=H(K,J|C_n)=$ $=H(K|C_n)+H(J|K,C_n)$; as $H(K^*|C_n)$ is known, it will be enough to compute $H(J|K,C_n)$ and then use:

$$(2) \qquad H(K|C_n) = H(K^*|C_n) - H(J|K,C_n)$$

Further: $H(K,M_n|C_n)=H(K|C_n)+H(M_n|K,C_n)=H(K|C_n)$ because M_n is a deterministic function of key and cryptogram; and also $H(K,M_n|C_n)=H(M_n|C_n)+$ $+H(K|M_n,C_n)$. By comparision (cf. /2/):

$$(3) \qquad H(M_n|C_n) = H(K|C_n) - H(K|M_n,C_n)$$

Now we deal directly with $H(J|K,C_n)$ and $H(K|M_n,C_n)$

Theorem 1.

$$H(J|K,C_n) = H(J) = \sum \log t_i!$$

Proof. Assume (k,j) and (k,i) are two extended keys for the s.s.c. (U, C) with the same actual key k. With respect to each other these keys only scramble equiprobable homophones relative to the same message-alphabet letter. Therefore, for any cryptogram $\underline{c}$, $\text{Prob}\{C_n=\underline{c}|K^*=(k,j)\}=$ $\text{Prob}\{C_n=\underline{c}|K^*=(k,i)\}$. This means that C_n and J are conditionally independent given K, and therefore $H(J|K,C_n)=H(J|K)$. But we already know that J and K are independent, so $H(J|K)=H(J)$. To complete the theorem, recall that J is a uniform r.v. with $t_1!t_2!\ldots t_s!$ values. QED

In theorem 2 the r-summation is extended over all exact types T_r in U^n and $h_i=h_i(T_r)$ is the number of distinct letters a_{ij} which do not occur in the sequences of T_r, $1\leq i\leq s$, $0\leq h_i\leq t_i$, $\sum h_i\leq t-1$.

Theorem 2.

$$H(K|M_n,C_n) = \sum_r P^{*n}(T_r) \left[\log(\sum h_i)! - \sum \log h_i!\right]$$

The non-zero terms in the summation are those for which at least two h_i's are positive, that is, at least two unequivalent homophones are

missing.

Proof. Assume that a couple message-cryptogram, $\underline{m}$, $\underline{c}$, is given of positive joint probability. Let us try to reconstruct the key, which is a sequence in A^t: e.g., if a_i and c_i are letters in the same position in the given couple, a_i is the j-th component of the key. However, gaps might be left because letter a_i might not occur, or it might occur in correspondence to less than t_i distinct cryptogram letters. If h_i denotes the number of times letter a_i is missing in the partially reconstructed key, the number of possible keys left is $(\sum h_i)!/h_1!h_2!\ldots h_s!$. Because of symmetry each such key has the same conditional probability, and so

$$H(K|M_n=\underline{m},\ C_n=\underline{c}) = \log(\sum h_i)! - \sum \log h_i!$$

Note that the integers h_i can be computed directly from the extended sequence $\underline{u}$ output by the extended source, h_i being simply the number of distinct letters a_{ij} which do not occur in $\underline{u}$. Note also that the set of $\underline{u}$-sequences with given integers h_i is a union of exact types. Therefore, grouping together $\underline{u}$-sequences in the same type:

$$H(K|M_n,C_n) = \sum_r P^{*n}(T_r)\left[\log(\sum h_i)! - \sum \log h_i!\right]$$

Clearly the quantity inside square brackets is zero only when at most one h_i is positive. This proves the last statement in the theorem. QED

Note that the key-appearance equivocation is zero only for t=2, and then the homophonic cipher is also a s.s.c. (s=t=2).

Now (1), (2) and (3), together with the two theorems, give the exact values of the equivocations $H(K|C_n)$, $H(K|M_n,C_n)$ and $H(M_n|C_n)$.

3. Asymptotic results.

It will be shown now that the key-appearance equivocation $H(K|M_n,C_n)$ becomes negligible with increasing message length n. Therefore for large n's both $H(K|C_n)$ and $H(M_n|C_n)$ are approximately equal to the constant term

$$\log A - \sum \log t_i!$$

("unremovable uncertainty"); cf. also the observations below formula (1). Note that the factors d_j which appear in the definition of A (cf. again (1)) are made up summing one or more t_i's bacause equivalent

homophones have all the same probability. So, as it should be, the unremovable uncertainty is non-negative. It is zero when only equivalent homophones are equiprobable (the numbers d_j and the numbers t_i are the same up to their order).

Now we investigate the behaviour of $H(K|M_n,C_n)$ as a function of n. To extend the validity of theorem 3 below to the case t=2, when $H(K|M_n,C_n)$ is zero, we adopt the (natural) convention that a term of the form $\exp\{n[-\infty+\varepsilon_n]\}$, $\lim_n \varepsilon_n=0$, means zero. We set $D=D(P,t_1,t_2,\ldots,t_s)=\min_{1\le i<f\le s}(p_i/t_i+p_f/t_f)\le 1$ (D=1 if and only if s=t=2).

Theorem 3.

$$H(K|M_n,C_n) = \exp\{n[\log(1-D)+\varepsilon_n]\},\ \lim_n \varepsilon_n=0$$

Proof. Take $t\ge 3$. We start with the obvious bounds:

$$\log 2 \sum_r P^{*n}(T_r) \le H(K|M_n,C_n) \le \log(t-1)! \sum_r P^{*n}(T_r)$$

the summations being restricted to types which correspond to non-zero terms in the summation of theorem 2. Denote by $M(i,j;f,g)$ the set of U^n-sequences such that a_{ij} and a_{fg} are missing in them; $1\le i<f\le s$, $1\le j\le t_i$, $1\le g\le t_j$; $M(i,j;f,g)$ is a union of exact types. One has:

$$\sum_r P^{*n}(T_r) = P^{*n}(\bigcup_r T_r) = P^{*n}(\bigcup M(i,j;f,g));$$

the sets in the latter union, which is not disjoint, are no more than $\binom{s}{2}[(t-1)!]^2$. One has also:

$$P^{*n}(M(i,1;f,1)) = P^{*n}(M(i,j;f,g)) = (1-p^*_{i1}-p^*_{f1})^n = (1-p_i/t_i-p_f/t_f)^n$$

Assume that D is achieved, say, for i=1, f=2. Then:

$$P^{*n}(M(1,1;2,1) = (1-D)^n \ge P^{*n}(M(i,j;f,g))$$

and the bounds for $H(K|M_n,C_n)$ can be relaxed to:

$$\log 2\ (1-D)^n \le H(K|M_n,C_n) \le \binom{s}{2}[(t-1)!]^2 \log(t-1)!\,(1-D)^n$$

This ends the proof. QED

Observe that the proof of the theorem implicitly gives asymptotically tight bounds for ε_n which are independent of $P(t\ge 3)$:

$$n^{-1}\log\log 2 \le \varepsilon_n \le n^{-1}\log\{\binom{s}{2}[(t-1)!]^2 \log(t-1)!\}$$

The parameter D which appears in theorem 3 does not coincide with the corresponding parameter obtained by Dunham /2/ for the key appearance equivocation $H(K^*|U_n,C_n)$ of the s.s.c. (U, C). He proved that $H(K^*|U_n,C_n)=\exp\{n[\log(1-\hat{D})+\hat{\varepsilon}_n]\}$, $\lim_n \hat{\varepsilon}_n=0$, where $\hat{D}$ is the sum of the two smallest components of P*; since these components may be relative to equivalent homophones, one has $\hat{D}\le D$.

The asymptotic behaviour of $H(K^*|C_n)$ is well-known (cf. /1,7/), and so we have all we need. We shall write down explicitly the asymptotic formula for $H(M_n|C_n)$, the most complex (and in a way the most relevant) of the three equivocations:

$$H(M_n|C_n) = H(K|C_n) - H(K|M_n,C_n) = \log A - \sum \log t_i!$$
$$+ \exp\{n[\log(1-B)+\delta_n]\} - \exp\{n[\log(1-D)+\varepsilon_n]\},\ \lim_n \delta_n = \lim_n \varepsilon_n = 0$$

$B=B(P,t_1,t_2,\ldots,t_s)$ is defined as $\min(\sqrt{p_i/t_i}-\sqrt{p_j/t_j})^2$, the minimum being taken over all distinct P*-probabilities, $1\leq i,j\leq s$, $p_i/t_i \neq p_j/t_j$; B is set equal to 1 for P* uniform , so that the corresponding exponential term becomes zero. Of course, if B<D, one can write the message equivocation as:

$$H(M_n|C_n) = \log A - \sum \log t_i! + \exp\{n[\log(1-B)+\delta'_n]\},\ \lim_n \delta'_n = 0,$$

while, if B>D, one has instead:

$$H(M_n|C_n) = \log A - \sum \log t_i! - \exp\{n[\log(1-D)+\varepsilon'_n]\},\ \lim_n \varepsilon'_n = 0.$$

4. Final remarks.

At the beginning of section 3 it has already been pointed out that, for large message lengths n, both the key and the message equivocation are approximately equal to the "unremovable uncertainty" $\log A - \sum \log t_i! \geq 0$. The condition for the unremovable uncertainty to be zero is that only equivalent homophones are equiprobable. An advantageous situation is found instead when P* is uniform (all the homophones are equiprobable); then the homophonic cipher is said to be matched (cf. /5/) and the unremovable uncertainty equals $\log t! - \sum \log(tp_i)!$, tp_i integers. In principle, when the components of P, the p.d. of the message source, are rational, one can always achieve P* uniform for a sufficiently large cryptogram-alphabet size t; however, alphabet extension runs counter to complexity requirements (it also leads to the growth of the term $\sum \log t_i!$). Once a threshold T>s is given, always assuming that the cipher will be used for a long time, one should judiciously choose the parameters $t,t_1,t_2,\ldots,t_s$, $s\leq t\leq T$, in order to achieve a large unremovable uncertainty. Were it not so, the performances of the homophonic cipher might even be worse than those of the s.s.c. (A, A) for the same

message source, for example when equal probabilities p_i and p_j are split to give distinct probabilities p_i/t_i and p_j/t_j, so that letters a_i and a_j become statistically distinguishable only in the case of the homophonic cipher.

Assume a cipher system (K,M_n,C_n) is given. We call two keys k and h indistinguishable when $T_k^{-1}(\underline{c})=T_h^{-1}(\underline{c})$ for all cryptograms $\underline{c}$, of any length ($T_k^{-1}(\cdot)$ denotes the cryptogram-to-message transformation determined by key k; note that in the case of a homophonic cipher the message-to-cryptogram transformation $T_k(\cdot)$ is not deterministically defined). The spy (and also the authorized receiver, for that) is interested only in the equivalence class of indistinguishable keys to which k belongs, rather than in k itself. Sometimes the extended key, K*, has been misinterpreted as the "true" key of a (strictly) homophonic cipher. If one neglects the fact that distinct extended keys with the same actual key are indistinguishable, one is lead to give over-optimistic evaluations of the cipher's performances. In particular, the negative term $-\sum \log t_i!$, which appears both in key and message equivocation, is ignored. Our description of the various types of "keys" in terms of suitable exact types makes it transparent why the "true" key of a homophonic cipher is precisely the actual key, K. Distinct actual keys are always distinguishable.

References.

/1/ R.J. Blom, "Bounds on key equivocation for simple substitution ciphers",IEEE Trans. Inform. Theory, vol. IT-25, pp. 8-18,Jan 1979.

/2/ J.G. Dunham, "Bounds on message equivocation for simple substitution ciphers", IEEE Trans. Inform. Theory, vol. IT-26, pp. 522-527, Sept. 1980

/3/ A. Sgarro, " Error probabilities for simple substitution ciphers", IEEE Trans. Inform. Theory, vol. IT-29, pp. 190-198, March 1983

/4/ A. Sgarro, "Exponential-type parameters and substitution ciphers"; a preliminary version called "Remarks on substitution ciphers" has been presented at the 1983 IEEE ISIT held at St. Jovite (Quebec)

/5/ J.G. Dunham, "Substitution and transposition ciphers", submitted

/6/ I. Csiszar and J. Körner, Information theory: Coding theorems for discrete memoryless systems. New York: Academic, 1981

/7/ A. Sgarro, "Simple substitution ciphers", in Secure digital communications, ed. by G. Longo. CISM Courses and Lectures No. 279, Springer-Verlag, Wien-New York, pp. 61-77, 1983