A Genetic Algorithm for Ciphertext-Only Attack in Cryptanalysis

Feng-Tse Lin
Dept. of Applied Mathematics
Chinese Culture University
Huakang,Taipei, Taiwan
E-mail : ftlin@ccu010.pccu.edu.tw

Cheng-Yan Kao
Dept. of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan

ABSTRACT

Cryptology is the science and study of systems for secret communications. It consists of two complementary fields of study : cryptography and cryptanalysis. In this paper, we propose a cryptanalysis method based on genetic algorithms to break the Vernam cipher. The proposed approach is a ciphertext-only attack in which we don't know any plaintext; the only thing we have to know is that the plaintext is an English document. Let $M = m_1, m_2, \ldots$ denote a plaintext bit stream and $K = k_1, k_2, \ldots$ a key bit stream. The Vernam cipher generates a ciphertext bit stream $C = E_k(M) = c_1, c_2, \ldots$, where $c_i = (m_i + k_i) \bmod p$, p is a base. In our work, we first tried to find out the key stream $K = k_1, k_2, \ldots$ from an intercepted ciphertext C by genetic algorithms and then use them to break the cipher.

1. INTRODUCTION

Cryptology is the science and study of systems for secret communications. It consists of two complementary fields of study : cryptography, the design of secret communications systems, and cryptanalysis, the study of ways to compromise of secret communications systems [1]. Cryptology primarily has been applied in military and diplomatic communications systems, but other significant applications are becoming apparent. As computer networks gain popularity, more and more sensitive information is being transmitted over channel where eavesdropping and message interception are possible. To keep such sensitive information secure, we need mechanisms to allow a user to protect data transferred over the network. A cipher is a secret method of writing, whereby plaintext is transformed into ciphertext. The process of transforming plaintext into ciphertext is called encryption; the reverse process of transforming ciphertext into plaintext is called decryption. Both encryption and decryption are controlled by cryptographic key parameters.

Cryptanalysis is the science and study of methods of breaking ciphers. It is assumed that the ciphertext is sent over insecure communications lines and is available to the cryptanalyst. His aim is to recover the plaintext from the ciphertext without knowing the key parameters. A cipher is breakable if it is possible to determine the plaintext or key parameters from the ciphertext. There are three basic methods of attack : ciphertext-only, known-plaintext, and chosen-plaintext [1]. Under a ciphertext-only attack, a cryptanalyst must determine the key solely from intercepted ciphertext, though the method of encryption and certain probable words may be known. On the other hand, a known-plaintext attack requires a substantial amount of plaintext and ciphertext be known, and a chosen-plaintext attack is able to acquire the ciphertext corresponding to selected plaintext. Simmons [5] classifies cryptosystems as symmetric (one-key) and asymmetric (two-key). In one-key cryptosystems the enciphering and deciphering keys are the same. One-key systems provide an excellent way of enciphering users' private files and also provide an excellent way of protecting information transmitted over computer networks. The concept of two-key cryptosystems was introduced by Diffie and Hellman in 1976 [2]. They proposed a new method of encryption called public-key encryption which have also introduced a fourth kind of attack : a chosen-ciphertext attack. Nevertheless, we only concern with the ciphertext-only attack and the one-key cryptosystem in our current work.

A cipher using a nonrepeating random key stream is called a one-time pad. One-time pads are the only ciphers that achieve perfect secrecy because there is not enough information to break the cipher. The Vernam cipher was designed by Gilbert Vernam in 1917 [1] which is an implementation of one-time pad. However, using each key only once obviously leads to a severe key distribution problem, and the one-time pad is only useful for relatively short messages which are to be sent infrequently [4]. In this paper, we propose a cryptanalysis method based on genetic algorithms to break the Vernam cipher. We assume the Vernam cipher is an approximation to one-time pads for the

0-7803-2559-1/95 $4.00 © 1995 IEEE

practical purposes. The proposed approach is a ciphertext-only attack in which we don't know any plaintext; the only thing we have to know is that the plaintext is an English document. Let $M = m_1, m_2, \ldots$ denote a plaintext bit stream and $K = k_1, k_2, \ldots$ a key bit stream. The Vernam cipher generates a ciphertext bit stream $C = E_k(M) = c_1, c_2, \ldots$,where $c_i = (m_i + k_i) \bmod p$, and p is a base. Since it is a one-key cryptosystem, the decrypted formula has the same key bit stream K as the encrypted formula, which is $M = D_k(C) = m_1, m_2, \ldots$, where $m_i = (c_i - k_i) \bmod p$. Figure 1.1 illustrates the enciphering and deciphering of data in the cryptographic system. The proposed approach will find out the key stream $K = k_1, k_2, \ldots$ from an intercepted ciphertext C and then use them to break the Vernam cipher.

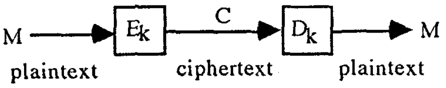

Fig. 1.1 Cryptographic system.

2. GENETIC ALGORITHMS

John Holland is the founder of the field of genetic algorithms(GAs). With the publication of *Adaptation in Natural and Artificial Systems* in 1975, Holland discussed the ability of simple bit-string representation to encode complicated structures and the power of simple transformations to improve such structures. GAs are stochastic adaptive algorithms that start with a population of randomly generated candidates and "evolve" towards better solutions by applying genetic operators such as crossover, mutation, and inversion, modeled on natural genetic inheritance and Darwinian survival-of-the-fitness principle. Over the past years, GAs have been applied to a variety of functional optimization problems, and have been shown to be highly effective in searching large, complex search space even in the presence of high-dimensionality, multimodality, and discontinuity. A GA is composed of a reproductive plan which provides an organizational framework for representing the pool of genotypes of a generation [3]. After the successful genotypes are selected from the last generation, a set of genetic operators are used in creating the offsprings of the next generation. Whenever some individuals exhibit better than average performance, the genetic information of these individuals will be reproduced more often. Consider the following outline of a GA shown in Fig.1.2. The GA simulates an evolutionary process with n individuals which represent n points in a large search space. From the view of engineering, GAs are an iterative process where each iteration has two steps, evaluate and generate. In the evaluation step, domain knowledge is used to determine the fitness of a candidate, a measure of its quality. Then, an evaluation function maps a candidate solution into the nonnegative real numbers. The generate step includes a selection operator and modification operators. The selection operator chooses individuals with a probability that corresponds to the relative fitness. Two chosen individuals, called the parents, produce children using the genetic operator crossover. The crossover operator exchanges substring of the codes of the parents at the same randomly determined point or points. However, it does not create any new genetic material in the knowledge base. The mutation operator, on the other hand, randomly changes a component in the structure introducing a new material into the knowledge base. Finally, the descendents replace some individuals in the population after the generation step is done.

```
1.  Initialize the parameters of the genetic algorithm.
2.  Initialize a population of chromosomes as the 0th generation.
3.  For generation := 1 to max_generation do the following steps;
4.  Calculate the fitness function for each chromosome.
5.  Calculate the summation of total fitness.
6.  While not done do the following steps
7.  Select parent chromosomes based on their fitness value.
8.  Apply the crossover and mutation operators to the parent chromosomes and produce a new generation.
9.  endwhile.
10. Replace the new generation as current generation.
11. endfor.
```

Fig. 1.2 The general outline of GAs

3. THE PROPOSED APPROACH

The proposed approach that based on genetic algorithms for breaking the Vernem cipher is now stated as follows. We first build a dictionary which consists of words that show up frequently in a general English document, such as the following words:

Dictionary = {this, it, has, have, do, does, as, been, what,
when, how, why, who, such, the, while, when, ...}.

Solutions for key stream of Vernam cipher are normally expressed as vectors of $K = \{k_1, k_2, \ldots k_r\}$. Thus each individual, or chromosome, in the population is encoded as an r-element solution vector. A gene is an element of the chromosome which denotes a key of the key stream. A good chromosome should have many correct keys that decipher the ciphertext to many correct words in the dictionary. Then, a good chromosome should have a high fitness value for reproduction. Once an individual has been selected for reproduction, an exact replica of the chromosome is made. This chromosome is then entered into a mating pool. Solutions in the population mate and bear offspring solutions in the next generation. After reproduction, a simple two-point crossover is performed. Crossover is the process by which two parent chromosomes recombine to create a new offspring chromosome. Over many generations the solutions in the population are improved until the best of the population is near optimal. Finally, we select the best chromosome which has the highest fitness value in the last population as the final solution. Then, we use the final solution as the key stream of the cryptosystem to decrypt an intercepted ciphertext *C* back into the plaintext *M*.

In the rest part of this paper, we assume the Vernam cipher is an approximation to one-time pads for the practical purposes. Let K' denote a chromosome of a key string k'_1, k'_2, ... k_r' and r is the length of the key stream. Again, as we have stated above, M = $m_1, m_2, \ldots m_n$ is a plaintext bit stream and C = $c_1, c_2, \ldots c_n$ is an intercepted ciphertext bit stream. The length of ciphertext is n. Let DIC be the decryption dictionary. A chromosome in the population is used to decode C into a certain kind of plaintext M' during the processing of the algorithm for calculating the fitness value. That is, M' = $D_k(C)$ =$m'_1, m'_2, \ldots,$ where $m'_i = (c_i - k_i) \bmod p$, and p is a well-known base. We define an array of Match[1.. r] called the match counters of a key string, where Match[i] is the number of times of the key string $k'_i\, k'_{i+1}$ whose decrypted string m'[i] m'[i+1] appears in the DIC. To calculate the fitness of a chromosome we simply square the value of each Match[i] and take their summation. That is, fitness = fitness + Match[i] * Match[i], for $1 \leq i \leq r$. The procedures of calculating fitness of a chromosome are stated as follows.

```
procedure fitness(K', C)
    initial: fitness = 0; Match[1.. r] = 0
    call decode(K', C, M')
    for i = 0 to n-1 do
        Match[ i mod r] = Match[ i mod r] +
                          find (m'[i], m'[i+1], DIC)
    endfor
    for i = 0 to r-1 do
        fitness = fitness + Match[i] * Match[i]
    endfor
    return fitness
end

procedure decode(K', C, M')
    for i = 1 to n do
        $m'_i = (c_i - k_i) \bmod p$.
    endfor
end

procedure find(m'[i], m'[i+1], DIC)
    if (m'[i], m'[i+1]) appear in DIC then
        return 1
      else return 0
end
```

For example, if a simple five-bit code is used, where the ith letter in the alphabet is represented by the binary representation of the number i (i.e. the letters A-Z corresponds to the numbers 1-26, respectively) and the space Δ is number 0. Assume that the plaintext M = {ATTACK AT DAWN}, the key string K = {22, 3, 15, 35, 28}, the length of string r = 5, and the base p = 25. The intercepted ciphertext C = {WWJKFHCPECADMX}. If one of chromosome is K' = {22, 3, 15, 16, 28} then the decrypted text M' = {ATTTCK AN DAWH}. The match counters are Match[1] = 3, Match[2] = 3, Match[3] = 2, Match[4] = 0, and Match[5] = 2. Thus, the calculation of the fitness of K' is $3^2 + 3^2 + 2^2 + 0^2 + 2^2 = 26$. On the other case, while we have chromosome K" = {12, 3, 15, 20, 28}, then the decrypted text M" = {KTTPCU AJ CAWU}. The match counters are Match[1] = 0, Match[2] = 3, Match[3] = 0, Match[4] = 0, and Match[5] = 0, and the fitness of K" is 9. The former is better than the latter.

The crossover operator we used in the algorithm is the usual crossover operator in the genetic algorithm literature[3]. Crossover takes two selected parents (chromosomes), splits the structures at the same two randomly determined crossing sites, and then creates two offsprings by swapping the middle portion of the structures. In other words, if the parents are A = $\{a_1, a_2, \ldots, a_n\}$ and B = $\{b_1, b_2, \ldots, b_n\}$, and the crossing sites are between b_i and b_{i+k}, then one child is $\{a_1, a_2, \ldots, a_{i-1}, b_i, \ldots, b_{i+k}, \ldots, a_n\}$ and the other is $\{b_1, b_2, \ldots, b_{i-1}, a_i, \ldots, a_{i+k}, \ldots, b_n\}$ (see Fig.1.3). Hence, the function of crossover is to generate rearrangements of coadapted groups of substructures from high performance structures. Furthermore, crossover has the main responsibility of mating high performance structures where the offsprings have schemata inherited from both parents.

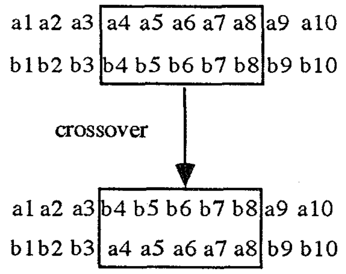

Fig. 1.3 Crossover operator.

4. EXPERIMENTAL RESULTS

In this section we provide experimental results to demonstrate the effectiveness of the proposed approach. We make the following assumptions in this experiment (1) r is unlimited, (2) m = 255 , and (3) the key parameters are randomly generated in [0, 255]. The plaintext are the first 107 words from the Declaration of Independence:

When, in the course of human events, it becomes necessary for one people to dissolve the political bands which have connected them with another, and to assume among the Powers of the earth the separate and equal station to which the Laws of Nature and of Nature's God entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the separation. We hold these truths to be self-evident; that all men are created equal, that they are endowed by their Creator with certain unalienable rights; that among these are Life, Liberty, and the pursuit of Happiness.

And, the intercepted ciphertext are:

íµÆâk0ûJ'ï£±Å®®¬¨∂¨q≈¨[∏Œâ£yoƒÿso□u7PôÃHìûn nr≠s
joû¬Å¢ry%}®¥sÕ≤∏Z6∫ßñ≠éÀIπóf≤∏ssolve%□œyÖs§Çz
√jçô´z°Åùìix'µòîí3Çßæàt•™â~üútedQÄ∂æâªÑ®Ã@L«ëâΩ¶
ΔàÄ!ªÑu5œyoΣÕu»ØçKbª'èör¤|äCU¥à°©≠_≠∏Xúé√q°∞ñz (
©Üt:∞¥†Ω≥}∏£]häÆB´∞§±náxÉkx©séto,ïÃ††mÉÜÀÑáïµ
ßágëßG[x◊¢rçÖ¬ëªEïñlöêŸÇ÷â+á6Vz}G§Ñ¥√ómrL vâj†}u±
q¶∏Ñ»¿respect4œzTúÃ¡\£÷pyæ°†°q«q.Ö™ôño®ÑÄ †quires
~òà∂V™¨äÅ1'≥â¡õhZùëÆ…√Æñ6tÜânéΔÃ»□÷CØ}≠àìÄ∞ùx
ñû>§£õøyñÅçΔñYΩgØ±»líj'Æx{ìo>{Æt≠\Øx¥se]ìò†£s'□ø
Σ[ñy»ªôèT≠™Å¥∞ßñâ7Œu¢Σd∞GçØΩ*¶reÑûáå¥u´óÇøìê¬–
3 "}vΣ□ èoö Σ)Δ—ïWï"≤uΩk´)°
Sõàf•zWjÑÉ†òÅx}'π≤Ç{ufÃèainI£
nalienable[ûßghts;C"ÆÄç¬"|∂∫ëïJÑÀeæª□≈àfsvÑkà{>Wãç}
Üå´=Öh∫°lî»™Çr£ã£∏≠œ6of{^ã~ö°¨ØΔèì)

The key streams of the Vernam cipher are :

59 77 73 27 63 16 103 48 42 33 59 76 97 69 57 77 58 67
71 81 86 70 59 80 89 28 66 11 79 95 98 14 1 11 2 11 48 48
88 40 49 57 11 91 1 13 58 83 99 5 12 57 79 14 65 0 0 5 23
57 66 83 94 68 83 58 54 74 66 39 61 34 102 41 69 40 70
78 79 0 0 0 0 0 0 5 11 103 20 101 3 53 22 17 79 1 42 56 63
90 90 32 47 47 32 88 93 77 47 49 42 19 26 70 72 35 84 66
59 27 16 58 57 0 0 0 49 12 78 89 28 7 68 27 52 100 32 76
102 35 26 73 62 97 22 84 1 90 22 17 21 91 10 79 86 90 2
83 66 40 43 1 78 102 33 51 82 103 20 37 35 5 69 17 87 55
58 63 62 82 56 40 38 94 81 60 79 36 6 98 8 53 30 15 26 61
79 48 92 65 28 68 62 61 7 28 74 34 0 70 63 47 80 2 103 5
15 10 4 64 4 32 0 0 0 12 30 100 55 61 5 99 18 99 31 103
73 84 48 20 71 34 65 39 13 23 99 45 0 40 101 97 35 87 37
38 48 76 76 47 101 13 100 36 4 20 22 15 11 25 39 63 22
64 90 35 1 13 76 2 33 5 51 81 85 80 0 13 65 85 31 90 76 0
0 0 0 0 0 0 0 20 91 11 52 40 100 92 60 52 102 7 11 85 50
50 46 81 88 11 14 24 73 43 43 6 58 32 96 88 59 0 0 0 0 0 0
94 36 32 85 92 54 54 68 37 8 49 98 75 26 76 47 4 58 57 44
75 93 98 60 49 22 0 30 36 78 43 101 87 85 26 99 35 56 21
68 37 43 96 71 48 8 49 50 30 48 59 54 82 89 92 39 97 25
94 49 57 74 2 63 80 86 11 30 1 101 64 74 91 60 10 30 19
63 8 73 92 59 16 79 0 0 61 31 38 43 47 11 98 95 75 72 59
10 48 89 85 86 45 41 39 72 52 24 80 75 57 34 78 23 90 13
65 67 0 3 80 80 39 32 74 79 10 69 0 0 100 59 21 39 83 1
70 51 98 90 34 27 97 100 7 0 94 21 21 67 21 27 7 53 62 9
101 95 48 55 48 100 78 6 70 6 71 9 63 87 51 39 32 1 60 8
55 39 18 30 63 36 18 6 93 93 80 62 26 91 18 1 90 27 0 0 0
41 46 0 0 0 0 0 0 0 0 0 0 59 44 62 0 0 0 0 0 35 34 58 24 44
78 2 27 73 75 34 46 42 16 99 0 75 86 95 100 22 1 83 42 27
5 35 79 30 11 34 43 24 20 24 50 17 101 7 76 61 76 32 96
69 98 2 46 25 48 67 68 91 22 0 0 91 22 42 14 42 56 62 74
83 28 101 41

Next, we list some partial results as follows, which are generated by genetic algorithm.

(1) one of the results in the 10th generation :

dhen@3jp&Çvj4course;|r9i□èaneventsI+àz7cqråãqå"x|eÄ
ééiÖä(oor=tâe$□wÑtÑÇr:â□.dissolve the âçnwyã□en%□ É
zds6which isve/ÑrÜnected*yhem=ìäth5□
anotherL9mqs@ Åê(o{êáÉp>Ççong"the0]swers8Öp the
earth=the separate !and$-
ÅÄêeã3ÅÉation7ês.Å}á~}!theAgaws'of6XlÑ}tÑ
and&Üo1dezÉre's!eåÅ()Ötitle vmÖn7 a@d{fnqÄ respect
to(çk~,}ãp{ions@of mankind1rtsãowÑÑ:uyh}&they
çÄÅîÑÑ declare0ínt@oÉuses2áz~yw.nmpel*âàám to1ãá
v7|ÜÉsration.=We hold □Çsåi.èÇçyyÄ6Åã)sr self-evident
;-{Öqå all2çpx/mÜg%ÖÅÅfied*eâlç, that àtsëAare
endowed@qã%åheir/Kã}Äwëì with(ciãñcÇçBunalienable
@ÅÉgrîs;: Öhat5dyëåá these=□tx'km|rM"Liberty,
and,ëhe6pursuit0áÉ+efÄpiness.

(2) one of the results in the 30th generation :

|Ø≠—dLîvzw¬∂ courseÄΔf□≥íóênBeventsoaΩõ
becomesfwrfessary forC∞òû óy~xú□>ñ°
dissolveaªwpCpå°q□ò©sø bands6xëôÖäG¢¶ìî
connectedUähem withX* another/Vcöp&™ %§°sume&â†«Öµ
theHñ†æã¡ò of#æ«{wìù∂ñyIthe\ƒqѵÉoíü andP'
ÅÑãalS¡™ø≤…¡– toD¿•ñÉæKŒ ãsäaws of!Åjture and of*èã—
èëp`∏xr óh≥ntitle them,vaR∏ecent|°q•√yƒƒYëÕ
thew≠≈□nions of=~ankindcê°ïÜ ŒíÆP÷™at@üûey
ñhould&ÑπÆlare;ÉãΣ causes which[«ìÃélK≠{ël É†oÄyîú
separation.?ÇvK{ø∂zU w°êÆ° truthsQtÃ be~self-
evident;M¶Àat all menw±§ñV¬£µated%ã°"¬∫ÑoÄ∞at
ö°g"[îwπdendowedl∂◊<Üê°øÀ]ñreatorEß√«âfíÜµ©ginN'ƒâÖw
ñ… lpùÜJÑΩåèóæ;-)

(3) one of the esults in the 50th generation :

When;:èÉ9ëèe@làçô~k.éf-
èöîin:f{ëxàõL%k{6kwyô}fâJtÄ äãúyaryFfor }|Ñ:&
òmèyàj?}u, âëïsolve3the:ôëlåìíåqò2
bands"óÖíÄh2nÑîe6ÜÅnnected<them.ëith0□ lnother;8át
Ñ0ëu assume8fñäng:ìëi&bwwersDof:|}áGhsrthEâqe:
separate5ãé}L~}ual$station#yóLwhich.|ugHVaws3Äu%W
Ñture:cñs$í}8Wråàre's9God ÖìäÖüsn% them-HnGiqÇent7
{ããâ{□v6{à'ñ{n5ãñinions/ofEsankindDrequiresCëèlà:òiq•
□ êåÇÅym declare the causes&òtmfh8□mpel-
them>êxFÉhe!ûlparation.GWeJhold
these$îâwuíâFôo#ÉuIÇãsÕDkúkfåuéY+çhkç9all-
Üen9Ére)yëÅwuÅd-wÅ°hÜ@ that íÄçá
are>endowed,lÇHyheir&Nàeator8çvÖÖ
certain!~talienable'ôtm{□éBL,
ÖnäùEàçóòã0ÅíeèäAtre6^ÑÖx57WÅlzèzl,GqndFÇÉlKá~zà
†àÑFofBHappiness.

(4) one of the results in the 50th generation :

Wien, jo!the course!pf!iunan fveous-!it cecpnfs!nfcettbry
fos one! qepqle uo!dissolve the political bands
which!iave!connected uien with □ bnother- and!uo!attvme
bnong the Qpxfst!og uhe earth the sfpbsauf and frubl
ttation to!xijdi!tif!Maws!pf Oauuse!and pg Obture(t
Hpe!fouitle□ them-!b decent respect to the oqinipns
pf!nankind rfrujsft that they tiould eeclare!tie dauses
which!jmpel uhem to uhf tepbration.!We ipld thesf
truths!uo ce self-evident; uiat alm!men bsf!crfbuee
fqubl,!thbt tiez are endowed!by tieir!Dsebtor xiuh

```
certain!unalienable!rights; □ tibt!ampng
these!are!Ljge,!Licerty- ane thf!qvstuit of!Happiness.
```

(5) one of the results in the 80th generation :

```
When, in the_courserFf humanhevents□ it becomes
necessary for one ≥ people to dissolve the political bands
which have connected“9/emhwith' another,nand to
assume˜J°≤ngÂGhe Powers of□Uhe earth the
separateKandx□ Uqual station to^!íich the
Laws<ØfOñaturepand?of Nature'sCGod entitle z©™mÄ=p
decent respect$toNthenopinions of mankind requires
thatmthey should declareXthe causes+∞hichO%mpel them
to(Ñ□e separation. We”hold these*±ruths toÃbe self-
evident; that all menDjre created equal˜ that they are
endowed byhtheir«Xreator with”certain unalienable
rights;+ □±Ut amongg$hese are.LifeÖXLiberty, and the
pursuit ofoHappiness.
```

5. CONCLUSIONS

We have presented a genetic algorithm method for breaking the Vernam cipher. The proposed approach is a ciphertext-only attack in which we don't know any plaintext in advance. It will find out the key stream from an intercepted ciphertext and use them to break the Vernam cipher.

REFERENCES

[1] Denning, D. E., *Cryptography and Data Security*, Addison-Wesley Publishing Company, Reading Mass., 1982.

[2] Diffie, W. and Hellman, M. E., "New Directions in Cryptography", *IEEE Transactions on nformation Theory*, Vol. 22, No. 6, pp. 644 - 654, 1976.

[3] Goldberg, D. E., *Genetic algorithms: In search, Optimization and Machine Learning*, Addison-Wesley Publishing Company, Reading Mass., 1989.

[4] Lempel, A.," Cryptology in Transition", *Computing Surveys*, Vol. 11, NO. 4, pp. 286 - 303, 1979.

[5] Simmons, G. J.,"Symmetric and Asymmetric Encryption", *Computing Surveys*, Vol. 11, NO. 4, pp. 304 - 330, 1979.