Strength of Crypto-Semantic System of Tabular Data Protection

Hazem (Moh'd Said) Abdel Majid Hatamleh
Computer Department, Al-Balqa' Applied University,
Faculty of Ajloun, Jordan

Roba mohmoud ali aloglah
Computer Department, Al-Balqa' Applied University,
Faculty of Ajloun, Jordan

Hassan Mohammad
AL-Ahliyya amman university
Department of Engineering, Amman Private University,
Amman, Jordan

Saleh Ebrahim Alomar
AL-Ahliyya amman university
Department of Engineering, Amman Private University,
Amman, Jordan

***Abstract*—The strength of the crypto-semantic method (CSM) of text data protection based on the use of lexicographical systems in the form of applied linguistic corpora within the formally defined restrictions of selected spheres of applied uses has been analyzed. The levels of cryptographic strength provided by the crypto-semantic method of data protection with due regard of a cryptanalyst's resource capabilities are determined. The conditions under which the CSM provides absolute guarantee of text data protection from confidentiality compromise are determined.**

Keyword—cipher key; cryptographic; data protection; crypto-semantic; lexicographical systems

I. Introduction

In [1], a text data protection method entitled "crypto-semantic method" (CSM) is suggested. The method is based on the use of lexicographical systems in the form of applied linguistic corpora within the formally defined restrictions of selected spheres of applied uses [2,3]. The CSM provides absolute guarantee of text data protection from confidentiality compromise even under the conditions when a sufficiently large number of encrypted information samples (demonstrably larger than the volume of password information) is available to the cryptanalyst. However, in [1] no cryptanalysis as to the CSM's strength has been made. No conditions and restrictions under which the use of the CSM is expedient have been defined. No correspondent formal foundations and proofs have either been provided. The present article aims to eliminate this deficiency.

In this paper, to define the levels of cryptographic strength which the crypto-semantic method of text data protection is capable of providing, with due account of a cryptanalyst's resource capabilities. To define the conditions under which the CSM provides absolute guarantee of text data protection against confidentiality compromise.

II. Analysis of a Cryptanalyst's Possible Actions

The cryptanalysis of the CSM system of tabular data protection under different conditions of its practical use is presented below.

A. *Initial Conditions.*

- It is known to the cryptanalyst that the secure text exchange channel functions according to the model presented in figure 1. The flowchart and the performance features of the CSM data protection system implementing this model is dealt with in [1]. The concept of this system is based on the synchronization of pseudo random sequence generators (PRSG) located on the transmitting and receiving sides of the secure exchange channel with the help of a known ciphering key [4-6].
- The text information to be encrypted is presented in a table of an arbitrary type. The form of the table is predefined. No information other than that entered into the table is available.
- The implementers of the CSM protection system, including the application area thesaurus, identical to the implementers of the secure exchange parties are available to the cryptanalyst.

1) Attack model #1

Below, the strength of the CSM protection system under the conditions when at least one pair of corresponding samples of tabular data (i.e. a plain original sample of tabular data and a corresponding sample of encrypted data) are known to the cryptanalyst is analyzed. The aim of the attack is for the cryptanalyst to determine the secret keyword (password) which a priory is unknown. It is appropriate in this case to take as the strength index the criterion K_1 – the maximum possible number of the brute-force search variants of the ciphering key equal to the number of possible ciphering key values:

$$K_1 = a^k$$

where a is the basis of the key information alphabet and k is the ciphering key capacity. The index K1 under the given conditions characterizes the strength of the CSM protection system on a specified fixed level which can be explicitly ascertained by the cryptanalyst.

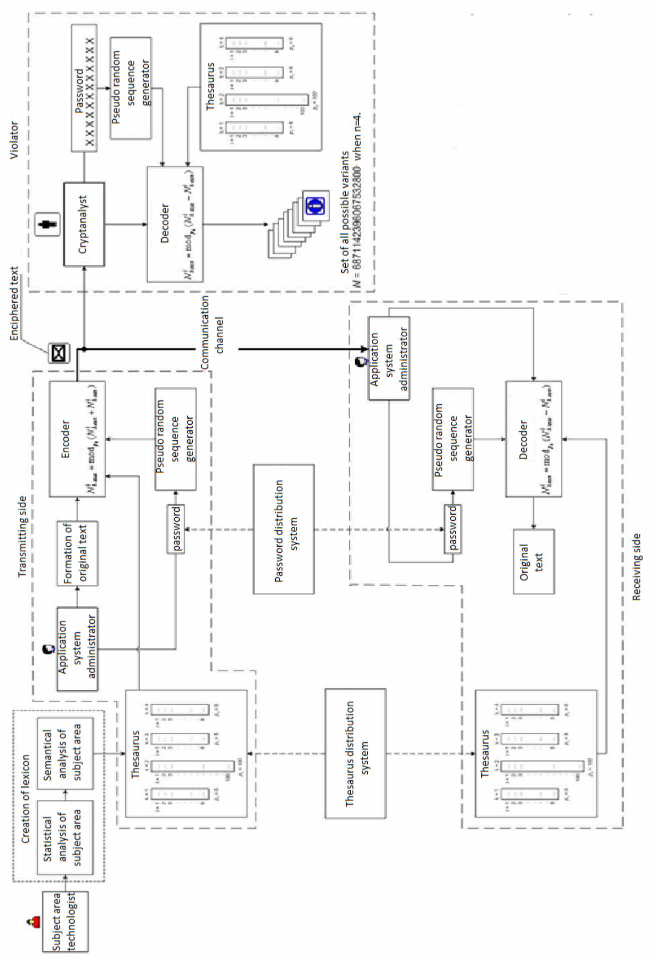
Violator
Password
XXXXXXXXXXXX
Pseudo random sequence generator
Thesaurus
Cryptanalyst
Decoder
Set of all possible variants
N = 6871142396067532800 when n=4.
Enciphered text
Communication channel
Application system administrator
Encoder
Pseudo random sequence generator
Decoder
Transmitting side
Formation of original text
password
Password distribution system
password
Original text
Receiving side
Application system administrator
Thesaurus
Thesaurus distribution system
Thesaurus
Creation of lexicon
Semantical analysis of subject area
Statistical analysis of subject area
Subject area technologist

Fig. 1. Functional model of the CSM system of tabular data protection

B. *Initial conditions:*

- the ciphering key is unknown, but at least one pair of samples – the original data and the corresponding encrypted tabular data – are known to the cryptanalyst.
- the secure exchange parties have not changed the password within the time span when these samples were received.

1) *Cryptanalyst's actions*

The cryptanalyst repeatedly attempts to decipher the encrypted sample of the tabular data, the original denotation of which is known to them, by brute force attack. In the process of deciphering, the cryptanalyst uses the implementers of the CSM tabular data protection system identical to the implementers of the secure exchange parties. The original denotation of the password is defined as the variant of the keyword with the use of which the corresponding known original sample of tabular data will be obtained as the result of deciphering.

2) *Conclusion on the attack model #1:*

- the cryptanalyst is able to identify the fact of the successful termination of the attack and, having implemented the attack model #1, to reliably determine the password.
- under these conditions the CSM system is unable to absolutely (according to Shannon [7]) guarantee protection. The CSM system's strength in these circumstances totally depends on the strength of the cryptographic algorithm used.

3) *Attack model #2*

Below, the strength of the CSM protection system under the conditions when no corresponding pairs of samples of original and encrypted tabular data are available is analysed. Here, the value of the distance of uniqueness does not meet the requirement of absolute protection guarantee (see [7]). The distance of uniqueness (or the point of uniqueness) is defined as such an approximate encrypted data size for which the sum of the real amount of information (entropy) in the corresponding plain data sample plus the keyword entropy equals the number of bits in the encrypted data sample. The distance of uniqueness is the cut-off criterion enabling to evaluate the minimum required volume of encrypted data samples sufficient for their brute-force deciphering. In the case when the analyst deciphers these data they are certain that they have obtained a reliable sample of the original data, as in this case only one reasonable way of their deciphering exists. The distance of uniqueness criterion serves not only as the measure of volume of intercepted encrypted data necessary for their deciphering, but also as the measure of volume of encrypted data samples necessary for the certainty in the uniqueness of the deciphering result obtained to exist. In this case it is considered that the volume of encrypted data available to the cryptanalyst exceeds the distance of uniqueness. Thus, a theoretical possibility of breaking the cipher exists.

4) *Initial conditions:*

- the absence of any corresponding pairs of original and encrypted information samples, i.e. cryptanalysis can be carried out only on the basis of the intercepted ciphertext;
- the implemented variant of the pseudo random sequence generators in the CSM system (see. [1,6]) provides the randomness of substitutions;
- the cryptanalyst is able to obtain the data on the statistical properties of the application area thesaurus to the extent enabling to construct a function of distribution of a priori probabilities of occurrence of semantic units of the predefined table form on the receiving end of the CSM protection system encoder;
- the cryptanalyst is able to obtain the volumes of ciphertexts exceeding the distance of uniqueness;
- the condition of maintaining the distance of uniqueness is not met; sufficient volume of the intercepted encrypted tabular data samples (obtained within the time span when the ciphering key was not changed) is available to the cryptanalyst in order for them to come to valid statistical conclusions as to the probability of a specific semantic units of the predefined table form appearing.

5) *Cryptanalyst's actions*

a) Preparatory stage.

Preliminary collection of the information on the statistical properties of the secure exchange information:

b) the collection of a batch of tabular data samples from the defined thesaurus within the defined application area with the use of the defined table form;

c) the statistical analysis of the collected batch with the aim of constructing a function of a priori probabilities distribution of occurrence of semantic units on the receiving end of the encoder which are the secure exchange information within the defined table form. This function may be used as a reference for the comparison with a posteriori probabilities distribution in the frequency analysis of the intercepted encrypted tabular data samples.

C. *Attack stage.*

The cryptanalyst uses the implementers of the CSM tabular data protection system, carries out the enciphering/deciphering of all variants of the semantic units of the defined table form by brute force attack and forms batches of variants corresponding to the intercepted encrypted tabular data samples.

The obtained variant batches are used by the cryptanalyst to construct possible variants of the discrete function of distribution of a posteriori probabilities of occurrence of semantic units on the transmitting end of the decoder.

The constructed variants of the function of distribution of a posteriori probabilities of occurrence of semantic units on the transmitting end of the decoder are compared by the cryptanalyst with the reference function of distribution of a priori probabilities constructed at the preparatory stage in order to make the decision as to the most probable variants of the password.

The cryptanalyst makes the decision as to the most probable variants of the password corresponding to the variants of the function of distribution of a posteriori probabilities most similar to the reference function. The similarity criteria depend on the matter of the applied problem solved by the defined table form.

It is clear that under these circumstances the strength index may not be a fixture as it may occur that the occurrence of specific variants of semantic units of the defined table form at the receiving end of the encoder are not statistically related. Thus, the statistical analysis may turn out to be unsuccessful. Nevertheless, a probability to define the lower threshold of the CSM system strength exists.

In this case it is expedient to present the strength index as

$$K_2 = K_1 \times V \tag{2}$$

where K_1 is the strength index of the implemented cryptalgorithm and V is the total number of the brute-force search variants of the tabular data samples in the course of implementing attack model #2. It is clear that V=V1V2, where V1 is the number of variants of semantic units of the defined table form sent to the receiving end of the encoder and V2 is the number of the intercepted encrypted secure exchange information samples used in the course of the analysis.

1) *Conclusion on the attack model #2.*

a) The results of the frequency analysis of the predefined table form semantic units with the use of intercepted encrypted tabular data samples under certain circumstances may essentially enhance the probability of a correct detection of the password. However, the cryptanalyst is unable to identify the fact of the successful attack completion, and having implemented attack model #2, cannot guarantee the reliability of the password detection.

b) under these conditions the CSM system is unable to absolutely (according to Shannon [7]) guarantee tabular data protection.

c) The strength of the CSM protection system under these conditions even in the worst case, i.e. when attack model #2 has been successfully implemented, is estimated as V times higher than the strength of the cryptographic algorithm used.

Below, a graphic presentation of the dependence of the strength index K on the ciphering key capacity k in relation to the two attack models discussed above is given (see figure 2).

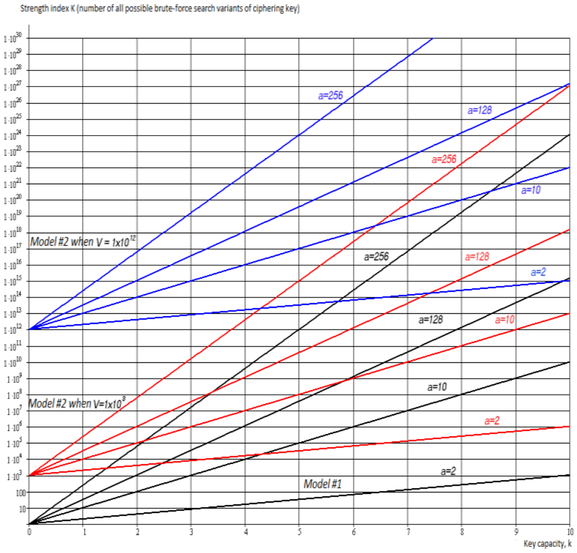

Fig. 2. Graph of strength index against ciphering key capacity

First, a trivial result: with the increase of the ciphering key length the protection system strength is enhanced. Second, in any case the strength of the CSM system against type 1 attacks is significantly lower than the strength of this system against type 2 attacks. Third, with the growth of the batch volume V the CSM protection system's cryptostrength is enhanced.

Below, the dependence of the strength index K on the a parameter, where a is the basis of the key information alphabet, is examined. It is seen in figure 1 that increasing the value of the a parameter will significantly increase the strength index K if k = const. For example, for attack model #1, where the key capacity k = 8, the strength index takes on the following values: K = 256 where a = 2, K = 1·108 where a = 10, K = 1,4064·1012 where a = 128 and K = 1,8447·1019 where a = 256.

The CSM system strength against type 1 attacks is significantly lower than this system's strength against type 2 attacks, and, with the growth of the batch volume V necessary for the analysis, the CSM protection system's cryptostrength is enhanced. In its turn, the need to increase V is conditioned by the demands for the increase of the statistical conclusions accuracy as to the character of the function of distribution of the semantic elements of the deciphered tabular data.

Use of the CSM protection system as a perfect secrecy system

As the basic protection effectiveness parameter we chose the so called amount of secrecy according to the terminology used in C. Shannon's works [7]. Also, the notion of the information protection system (IPS) entropy based on the use of key information is used. The IPS entropy is used as a measure of the amount of space of the password information keys. Assume the condition of maintaining the distance of uniqueness in this case in met. Insufficient volume of the intercepted encrypted tabular data samples (obtained within the time span when the ciphering key was not changed) is available to the cryptanalyst in order for them to come to valid statistical conclusions as to the probability of a specific semantic units of the predefined table form appearing.

In most symmetrical key systems, the distance of uniqueness is determined according to the following formula:

$$U = \frac{H(K)}{D} \tag{3}$$

where H(K) is the information protection system (IPS) entropy, K is the number of possible keys in the IPS and D is the redundancy of the language used for message display.

In its turn, the language redundancy D is calculated as

$$D = R - r \tag{4}$$

where R is the maximum entropy of stand-alone metasymbols, and r is the entropy of the language used for displaying the message M, calculated as

$$r = \frac{H(M)}{n} \tag{5}$$

where H(M) is the entropy of the message and n is the message length.

In this case enciphered samples with the total length less than the distance of uniqueness are used for encrypting messages. Thus, it is possible to provide a theoretically perfect protection, as under such circumstances the ambiguity of the cryptanalytical problem solving appears. If by means of the correct thesaurus synthesis one can provide almost equal probability of receiving each solution, under such circumstances the cryptanalyst find themselves in an ambiguous state, in particular they cannot make a valid decision, true on the basis of the deciphered messages.

Thus, in this case we stick to the condition that the volume of the tabular data encrypted by one key does not exceed the distance of uniqueness:

$$U = \frac{\log_2(K\ )}{D} \tag{6}$$

where U is the distance of uniqueness, K is the maximum possible number of the brute-force search variants of the ciphering key and D is the redundancy of the language used for displaying the semantic units of the predefined table form.

If condition (6) is met, in the case of an exhaustive search of the ciphering key the original sample of the transmitted data will appear on the transmitting end of the decoder not more than once.

Initial condition: the protection system meets the conditions of a perfect secrecy system, i.e. the cryptanalyst is unable to obtain volumes of data encrypted with one key exceeding the distance of uniqueness.

The definition of the strength index under these conditions loses any significance, since it is impossible to identify the moment of the successful attack completion. If the CSM system parameters meet the conditions of a perfect secrecy system, the tabular data protection is absolutely guaranteed. Neither a priori nor a posteriori data on the statistical properties of the secure exchange information can be used. Thus, modelling of any attacks under these circumstances loses its sense.

Also, under these circumstances an absolute protection guarantee is provided by the famous Mauborgne/Vernam scheme [4,5]. Below, the proofs that the CSM system has essential advantages over this scheme are presented.

Below, we plot the uniqueness distance as a function of the message length.

The IPS entropy H(K) is used as the measure of the amount of space of the keys K, namely:

$$H(K) = \log_2 K \tag{7}$$

where K is the number of possible keys in the IPS.

The language redundancy is calculated using formula (4). Consequently

$$R = \log_2 B \tag{8}$$

where B is the number of alphabet symbols calculated using the following formula:

$$B = \prod_{i=1}^{s} S_i \tag{9}$$

where s is the number of sublexicons in the selected tabular form thesaurus, Si is the number of words (or phrases) in the ith sublexicon of the thesaurus.

The entropy of the language r, with the help of which the message M is displayed, is calculated using formula (5). The entropy is measured in bits and equals

$$H(M) = \log_2 N \tag{10}$$

where N is the number of possible meanings of the message.

Thus, on the basis of (3), (8) and (5), we have:

$$.\quad U = \frac{H(K)}{\log_2(B) - \frac{H(M)}{n}} \tag{11}$$

On account of the perfect secrecy system properties, the number of keys K must equal N – the number of messages having the length of n. Thus, if $H(K) = H(M) = \log_2 N$, the following equation is possible:

$$U = \frac{\log_2 N}{\log_2(B) - \frac{\log_2 N}{n}} \tag{12}$$

It is now necessary to find the dependency of N – the number of possible message meanings – against n – the message length. When calculating N, it is worth keeping in mind that each table row (i.e. each letter in the message) occurs only once (i.e., letters do not repeat). In this case, the maximum possible message length equals the number of letters in the alphabet. Thus, the equation for calculating N – the number of possible message meanings at different n – is as follows:

$$N = \mathop{\Pi}_{n=1}^{n}\left[B-(n-1)\right] \tag{13}$$

In order to meet the condition of keeping the distance of uniqueness, it is necessary to correctly calculate the key capacity (length) in correlation with the message length:

$$k = \log_2 N \tag{14}$$

where k is the keyword capacity and N is the number of the possible meanings of a message having the length of n. With due account of (13), the dependency of the ciphering key length against the message length can be expressed in the following way:

$$k = \log_2 \mathop{\Pi}_{n=1}^{n}\left[B-(n-1)\right] , \tag{15}$$

where B is the number of the symbols of the alphabet of the language in which the message is presented.

Thus, where B = const and where $H(K) = H(M)$ (the condition of the perfect secrecy system), the following can be presented (see figure 3):

$$U(n) = \frac{\log_2 \mathop{\Pi}_{n=1}^{n}\left[B-(n-1)\right]}{\log_2(B) - \frac{\log_2 \mathop{\Pi}_{n=1}^{n}\left[B-(n-1)\right]}{n}} . \tag{16}$$

The dependency of the key entropy H(K) against the message length n can be presented in the following way (see figure 4 for the diagram):

$$H(K) = \log_2 \mathop{\Pi}_{n=1}^{n}\left[B-(n-1)\right] \tag{17}$$

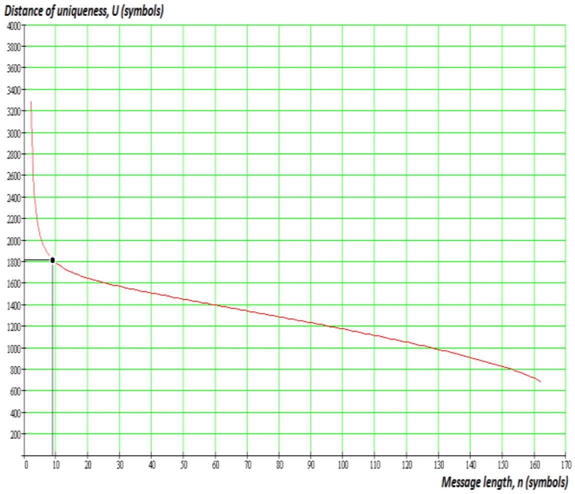

Fig. 3. Graph of the distance of uniqueness U against the message length n

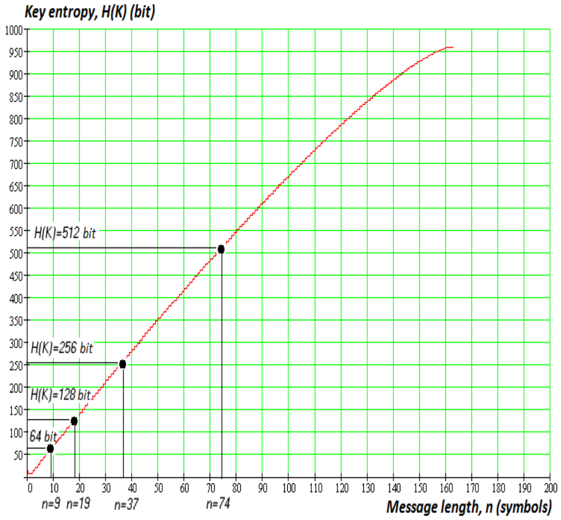

Fig. 4. Graph of the ciphering key entropy H(K) against the message length n

It is advisable to use the graphs in figures 3 and 4 for calculating the maximum possible number of communication sessions without changing the ciphering key while keeping the feature of the perfect secrecy system intact.

It is clear that the CSM protection system in the perfect secrecy system mode, with 64-bit key length with no change in the key, can ensure the absolute protection of 1800 transmitted language elements, while the one-time pad technique provides absolute protection only for 64 language elements. Thus, the

CSM protection system, as opposed to the famous Mauborgne/Vernam scheme, can provide absolute protection under the conditions that the volume of original (plain) text significantly exceeds the volume of key (password) information.

It is worth keeping in mind that the artificial language redundancy dealt with in the given example is negligibly small, which is uncharacteristic of natural languages. Artificial languages are characterized by relatively large value of the distance of uniqueness.

III. Main Results and Conclusions

The strength of the crypto-semantic method (CSM) of text data protection based on the use of lexicographical systems in the form of applied linguistic corpora has been analysed. The indices of cryptographic strength provided by the crypto-semantic method of text data protection with due regard of a cryptanalyst's resource capabilities are determined and the levels of cryptographic strength are introduced. The conditions under which the CSM provides absolute guarantee of text data protection against confidentiality compromise are determined.

If at least one pair of samples – original ones and corresponding samples of encrypted tabular data – are known to the cryptanalyst, the CSM system's strength in these circumstances totally depends on the strength of the cryptographic algorithm used.

If no corresponding pairs of original and encrypted information sample pairs are available (i.e. cryptanalysis is carried out only on the basis of the intercepted ciphertext), but the cryptanalyst is able to obtain the data on the statistical properties of the application area thesaurus and the volumes of ciphertexts exceeding the distance of uniqueness, the CSM protection system strength is estimated as V times higher than the strength of the cryptographic algorithm used, where V is the total number of the brute-force search variants of the tabular data samples.

If the CSM system meets the conditions of a perfect secrecy system, i.e. the cryptanalyst is unable to obtain volumes of data encrypted with one key exceeding the distance of uniqueness, protection is absolutely guaranteed. As opposed to the famous Mauborgne/Vernam scheme, the CSM system can provide absolute protection under the conditions that the volume of original (plain) text significantly exceeds the volume of key (password) information.

References

[1] The Art of Deception Kevin D. Mitnick 17 Oct 2003

[2] McEnery, Tony and Wilson, Andrew. Corpus Linguistics: An Introduction. 2nd Edition, Edinburgh: Edinburgh University Press, 2001.

[3] Schneier B., Applied Cryptography: Protocols, Algorithms, and Source Code in C, 2nd ed. New York // John Wiley and Sons, 1996.

[4] Математичні основи криптоаналізу [Текст]: навч. Посібник С.О.Сушко, Г.В. Кузнецов, Л.Я. Фомичова, А.В. Корабльов. – Дніпропетровськ (Україна): Національний гірничий університет, 2010. -465 с.

[5] Шеннон К.Э. «Теория связи в секретных системах». В кн.: Шеннон К.Э. Работы по теории информации и кибернетике. М.: Иностранная литература. 1963, с. 332-402, -829 с.

[6] A statistical Test Suite for the Validation of Random Number Generators and Pseudo Random Number Generators for Cryptographic Applications [Text]: NIST Special Publication 800-22 Rev1. – Gaithersburg, Maryland: NIST, 2008. – 153 p.

[7] Cryptanalysis Helen Fouche Gaines 01 Apr 1989

[8] Mike Meyers' CompTIA Security+ Certification Passport (Exam SY0-301)T. J. Samuelle 01 Jul 2011