Statistical Analysis of Text Encryption Software

Darko Brodić[1], Ivo R. Draganov[2] and Zoran N. Milivojević[3]

Abstract – **The paper describes the problem related to the efficacy of different text encryption programs. Many of the known programs uses similar techniques along with different keys to encrypt and decode the encrypted text. Still, many of them have different efficacy in the relation between output and input data, i.e. encrypted and original data. Also, some of them are more adjusted to different text document, which includes various languages and scripts as well as multilingual documents, too. This study statistically analyses aforementioned problems on the example of 5 different text encryption programs. The obtained results show meaningful differences between different encryption programs, which were under research process.**

Keywords – **Algorithm, Cipher, Cryptography, Encryption, Software, Statistical Analysis.**

I. INTRODUCTION

Encryption represents a process of transforming accessible data into an unreadable code that cannot be understood by normal means [1]. The encryption process uses a key as well as an algorithm to turn the accessible data into an encoded data. To decode encoded data or to convert encoded data to the original form, the ciphering algorithm and a secret key are needed. Hence, the encryption software includes the algorithm for encoding data in order that unauthorized users cannot access it [2]. These data can be accessed only if we have an appropriate key for decoding. Typically, the encryption process is used to protect our personal or business sensitive data. In this way, encryption software comprises different methods that may encrypt files, directories and drives to ensure data privacy. Currently, the application of this software is mainly in the area of archiving a large amounts of data or secure private communication over the Internet.

A whole group of different encryption software solution has been developed. Many of them differ according to their simplicity or complexity of included algorithms. Some of them includes less or more complex keys for coding and decoding. In the recent times, more and more solutions represent a freeware software ones.

The base of this study incorporates the research of different encryption programs and their implication on data that were subject to the encryption. In this way, the inputs to the encryption programs represent text documents given in different languages and scripts. The outputs represent encrypted data. The study statistically analyzes the output-input relationship. The conclusions are drawn according to the efficacy of the encryption level and suitability of using documents written in different languages and scripts.

The paper is organized in the following manner. Section 2 describes the encryption software under the research. Section 3 defines the experiment as well as the statistical measures used for comparison. Section 4 presents the results and gives the discussion. Section 5 draws conclusions.

II. TEXT ENCRYPTION PROGRAMS

In this study, the following encryption programs were under consideration and evaluation: (i) Arcanum Editor, (ii) BCTextEncoder, (iii) TheLetterEncrypter, (iv) Simple Text Encryptor, and (v) CryptX.

Arcanum Editor is a text encryption software. It is one of the best tools for encryption and decryption. It has multiple methods to encrypt your data. The various supported encryption algorithms are AES, Base64, Bytes, 1337 speak and Rot13. You can set an encryption password when using AES algorithm, while others simply encrypt/decrypt your text. It also has a Hash feature.

BCTextEncoder is a free text encryption software. It can encrypt typed text and text files. The encrypted text can be saved in a text file or copied from the clipboard. It has two types of encryption methods; one is password based encryption and the second is public key based encryption. It doesn't require installation, just download it and run it.

TheLetterEncrypter is a free text encryption software. You can encrypt or decrypt your text using any password of your choice. You can also import text file to encrypt and export the encrypted text to a file. It uses strong Rijndael Encryption algorithm to encrypt text. You can copy encrypted text to your email message to send. It requires no installation.

Simple Text Encryptor is a free and simple program that encrypts and decrypts text using 128-bit AES encryption method. It comes in two versions to download; one is an installer and other is stand alone zipped version. Zip version doesn't require installation and you can use it from any portable drive also. You can set or generate a random key to encrypt your text easily. It has very simple interface and is easy to use.

CryptX is a free text encryption software. You can encrypt and decrypt text files with it easily. It supports three types of encryption methods: Binary, Base64 and Base64 without password. Only Base64 is password supported encryption method. It saves the encrypted file in the normal text file. It has different tabs for encryption and decryption. It also has a feature of computing hash and calculate MD5 checksum.

[1]Darko Brodić is with the Technical Faculty in Bor, University of Belgrade, Vojske Jugoslavije 12, 19210 Bor, Serbia, E-mail: dbrodic@tf.bor.ac.rs

[2]Ivo R. Draganov is with the Faculty of Telecommunications at Technical University of Sofia, 8 Kl. Ohridski Blvd, Sofia 1000, Bulgaria, E-mail: idraganov@tu-sofia.bg

[3]Zoran N. Milivojević is with the College of Applied Technical Sciences, Aleksandra Medvedeva 20, 18000 Niš, Serbia, E-mail: Zoran.milivojevic@vtsnis.edu.rs

III. EXPERIMENT

The experiment is performed by encrypting a small database of text documents by all five chosen encryption programs. The database consists of 75 text documents, i.e. 25 English text documents, 25 Serbian text documents written in Latin script and 25 Serbian text documents written in Cyrillic script. The example of the text document samples is given in Fig. 1.

But in the ancient way of Himalayan travel, they're still 20 switchback roads and several torturous hours away.

(a)

U ostvarivanju ovog projekta, američki internet pretraživač se udružio sa šest fotografa, skijaša i planinara koji su sa kamerama u rancu prošli najpoznatijim stazama po letnjem i zimskom vremenu.

(b)

У остваривању овог пројекта, амерички интернет претраживач се удружио са шест фотографа, скијаша и планинара који су са камерама у ранцу прошли најпознатијим стазама по летњем и зимском времену.

(c)

Fig. 1. The example of the text document samples from the database: (a) English document, (b) Serbian Latin document, and (c) Serbian Cyrillic document

All documents included in the database represent a small text document. English documents consist of 75 to 176 characters, while Serbian Latin documents consist of 74 to 196 characters and Serbian Cyrillic documents consist of 73 to 196 characters. As a result, the input text is coded. The sample of encrypted documents is given in Fig. 2.

Lt6nrSTxksmvuc7I61ROMtnx9liVi8dRaq5P0Pe9ZKjnkag7D F2Q5ck5pOR9wPuk0btVctdu3wXvalvGoPr7ayp977aUAGT3 MYNon4H/K2gYEZNVcuuky+OK7Y2dVRioL4Xj5WGRqu RL/bdxz/AIqkyHUvftOQGr

(a)

XUAReMDMKUJQj9SWWV/N4w6z7t/EDBfxxcc15ma6Dzs Y0L6xoCGdidZL2a51vAbQHH7vjBHWFN5vjjqO9cxlTLj4q EkfrtXqvuogo/ZXfF2JI9UsBAWAfVqCPgnvlGfXSgziF7EA aw3ILrnKr9W1imUBUG+ksXO2qXK3JovPwU4Q8DohyeD n3h9F0ZZ7Wpf7Ydpg2UB3q/ZL66hqWz1ZnazJHKm1meD 7Nt91fiJCLWjP7G183kHuEYcHirkdzztn5vExI+KtWug=

(b)

0l6b8krM8vETLxnOcCgh7qGbQ2IezvTxyqLRgas/nrYTLxn OcCgh7vvWqxnvl5/1sPf4NIJYW6Wg0Mi7uRt8TRMvGc5w KCHuGQGX7v6e0kSZsFnr3+9/1eiR160rKsRLn3ihAvp3gE A/bQ95+zysB9VOScas7f22sPf4NIJYW6WfeKEC+neAQIQ HYoSNOIhFsPf4NIJYW6UTLxnOcCgh7hMvGc5wKCHuh AdihI04iEXSXpvySszy8dJem/JKzPLx/8jTYS9QtbU=

(c)

Fig. 2. Examples of the encrypted text document samples from the database: (a) encrypted English document, (b) encrypted Serbian Latin document, and (c) encrypted Serbian Cyrillic document

The output, which represents the coded text is statistically analyzed by typical statistical measures like mean, standard deviation and correlation coefficients. These measures are defined as follows, respectively:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \,, \tag{1}$$

$$s = SD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N} (x_i - \bar{x})^2} \,, \tag{2}$$

$$R = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2 \sum_{i=1}^{N} (y_i - \bar{y})^2}} \,. \tag{3}$$

The correlation coefficient R expresses the strength and direction of a linear relationship between two variables x (input) and y (output), where x_i is the value of the variable x for the instance i, $\bar{x}$ is the mean of x value, y_i the value of the variable y for the instance i, and $\bar{y}$ is the mean of y value [4]. Consequently, R measures the strength and direction of a linear relationship between two variables on a scatterplot [5]. It can receive the value from -1 to +1. The larger the absolute value of the coefficient, the stronger linear relationship between the variables (input and output). An absolute value of 1 indicates a perfect linear relationship. If R is positive, then the two variables tend to increase or decrease together. In contrast, if R is negative, then one variable increases as the other decreases.

IV. RESULTS AND DISCUSSION

Table I shows the statistical values obtained after encryption of Serbian Latin documents by all five encryption programs (Output1 - Arcanum Editor, Output2 - BCTextEncoder, Output3 - TheLetterEncrypter, Output4 - Simple Text Encryptor, Output5 – CryptX).

TABLE I
STATISTICAL MEASURES OF ENCRYPTED SERBIAN LATIN DOCUMENTS

	Output1	Output2	Output3	Output4	Output5
Min	108	236	108	160	108
Max	272	356	280	384	268
Mean	185.08	297.44	193.92	283.2	186.56
SD	43.69	28.78	43.60	60.21	41.80
R	0.9944	0.9768	0.9812	0.9814	0.9874

Table II shows the statistical values obtained after encryption of Serbian Cyrillic documents by all five encryption programs (Output1 - Arcanum Editor, Output2 -

BCTextEncoder, Output3 - TheLetterEncrypter , Output4 - Simple Text Encryptor, Output5 – CryptX).

TABLE II
STATISTICAL MEASURES OF ENCRYPTED SERBIAN CYRILLIC DOCUMENTS

	Output1	Output2	Output3	Output4	Output5
Min	176	284	192	160	128
Max	476	420	492	416	268
Mean	324.8	349.76	337.44	282.88	197.44
SD	79.57	34.56	80.14	63.79	39.76
R	0.9992	0.9892	0.9952	0.9902	0.9894

Table III shows the statistical values obtained after encryption of English documents by all five encryption programs (Output1 - Arcanum Editor, Output2 - BCTextEncoder, Output3 - TheLetterEncrypter, Output4 - Simple Text Encryptor, Output5 – CryptX).

TABLE III
STATISTICAL MEASURES OF ENCRYPTED ENGLISH DOCUMENTS

	Output1	Output2	Output3	Output4	Output5
Min	100	232	108	160	108
Max	236	324	256	384	248
Mean	158.12	273.92	170.68	248.32	168.48
SD	35.38	23.35	36.59	56.00	36.47
R	0.9992	0.9849	0.9835	0.9855	0.9957

From the results given in Tables I-III, it is quite clear that all encryption programs due to the high value of correlation coefficient *R* create linearly dependent output-input correlation. It means that the longer input document text will tend to create a longer coded output text. Furthermore, the analysis will be continued by comparison of the statistical analysis of documents written in different languages and/or scripts.

Fig. 3 shows the output-input relation in accordance to the number of characters obtained by the encryption program Arcanum Editor.

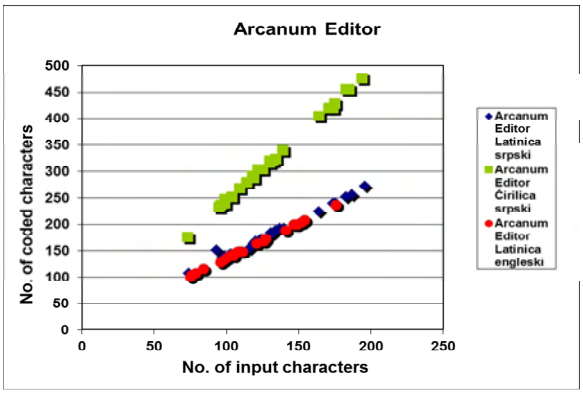

Fig. 3. Coded vs. input text obtained by Arcanum Editor

From Fig. 3, it is an obvious that the encryption of English and Serbian Latin documents has similar efficacy. However, the coding of Serbian Cyrillic documents creates larger coded documents. Hence, the program is more appropriate for the Latin based languages.

Fig. 4 shows the output-input relation in accordance to the number of characters obtained by the encryption program BCTextEncoder.

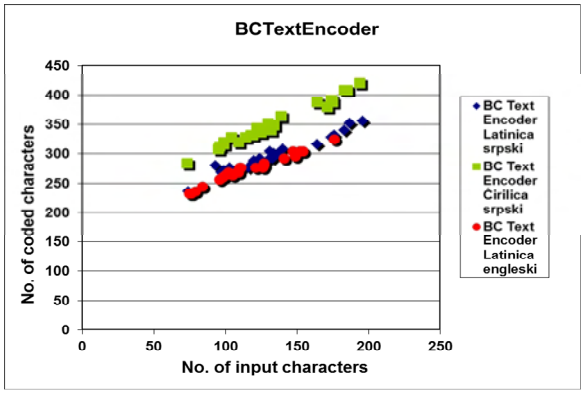

Fig. 4. Coded vs. input text obtained by BCTextEncoder

BCTextEncoder also depends on the type of the input scripts. Hence, English and Serbian Latin documents are similarly coded. As a consequence, Serbian Cyrillic documents are coded by creating larger documents.

Fig. 5 shows the output-input relation in accordance to the number of characters obtained by the encryption program TheLetterEncrypter.

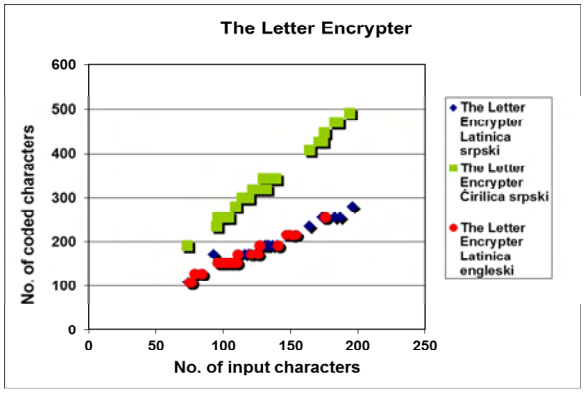

Fig. 5. Coded vs. input text obtained by TheLetterEncrypter

The situation with TheLetterEncrypter program is similar to previously analyzed programs.

Fig. 6 shows the output-input relation in accordance to the number of characters obtained by the encryption program Simple Text Encryptor.

Simple Text Encryptor similarly coded documents written by different languages/scripts. Hence, it has similar efficacy of encrypting document given by different languages and scripts.

Fig. 7 shows the output-input relation in accordance to the number of characters obtained by the encryption program CryptX.

CryptX program has similar way of encrypting documents in spite of their language or script. In this way, it is similar to the Simple Text Encryptor.

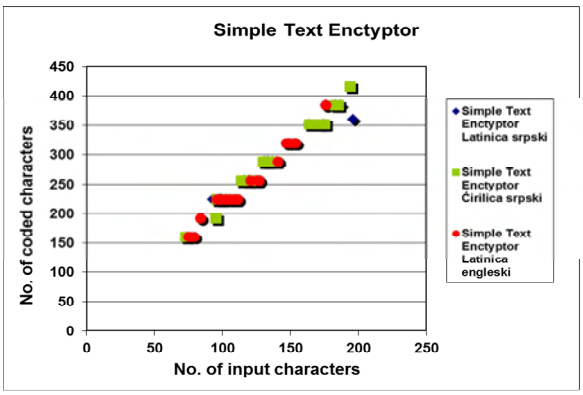

Fig. 6. Coded vs. input text obtained by Simple Text Encryptor

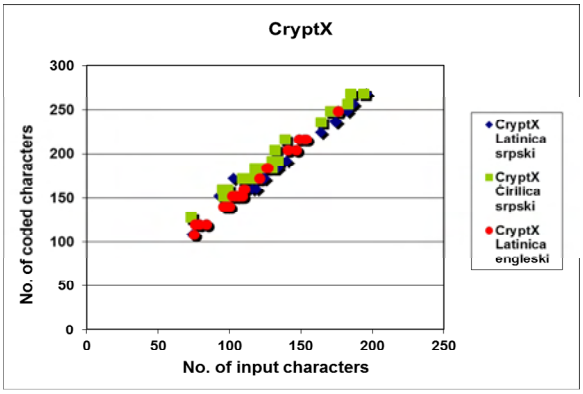

Fig. 7. Coded vs. input text obtained by CryptX

At the end, the comparison of different programs is established by giving their efficacy of encrypting documents in certain languages or scripts. Fig. 8 shows coded vs. input for English based documents.

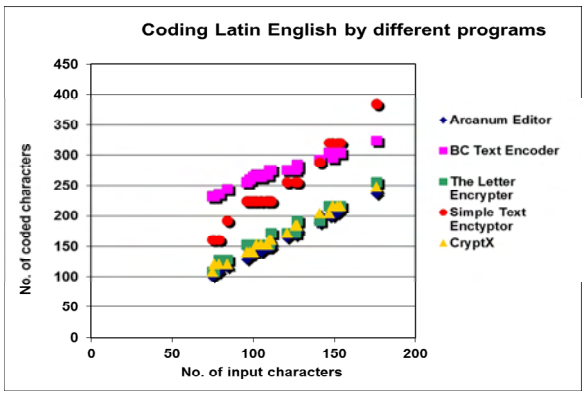

Fig. 8. Coded vs. input English text

Fig. 9 shows coded vs. input for Serbian Latin based documents, while Fig. 10 shows coded vs. input for Serbian Cyrillic based documents.

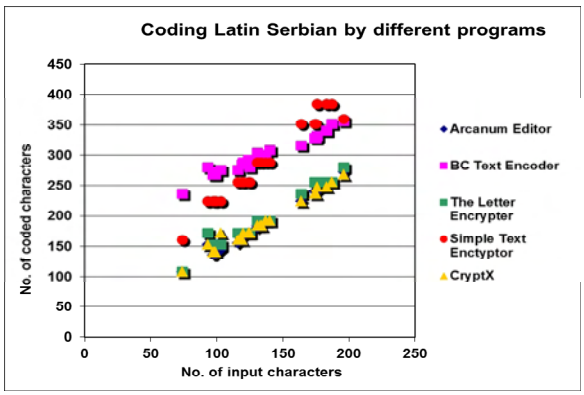

Fig. 9. Coded vs. input Serbian Latin text

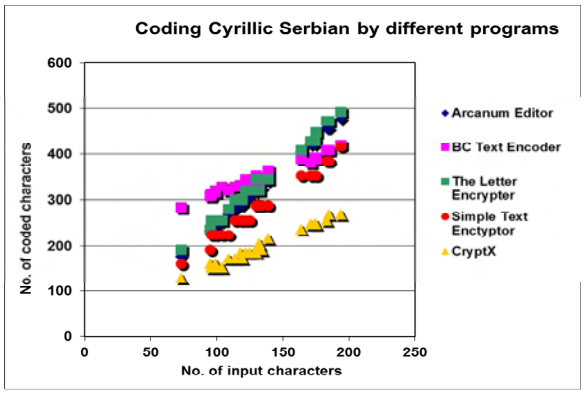

Fig. 10. Coded vs. input Serbian Cyrillic text

V. Conclusion

This paper analyzed different types of encryption programs. Their efficacy of coding was in the focus. The statistical analysis showed similarities as well as dissimilarities between their efficacy of coding. Many of the programs are primarily adjusted to the Latin based languages. However, some of them like Simple Text Encryptor and CryptX are almost completely independent of the type of document that they are coding. However, CryptX program created smaller coded documents achieving better efficacy.

Future research will include wider range of encryption programs as well as more complex database, which will include Chinese and Indian texts.

References

[1] A.J. Menezes, P. C. van Oorschot, S. A. Vanstone, *Handbook of Applied Cryptography*, Boca Raton, CRC Press, 1996.
[2] P C. Paar, J. Pelzl, *Understanding Cryptography*, New York, Springer, 2010.
[3] http://listoffreeware.com/list-best-free-text-encryption-software/
[4] J. Higgins, The Radical Statistician: A Beginners Guide to Unleashing the Power of Applied Statistics in The Real World (5th Ed.), Jim Higgins Publishing, 2006.
[5] Chi Yau, R Tutorial with Bayesian Statistics Using OpenBUGS, Kindle Book, 2015.