Protection of critical information in external data stores using the dissection-separation method

George N Shurkhovetsky
Irkutsk State Transport University, Chernyshevsky str., 15, Irkutsk, 664074, Irkutsk region, Siberian Federal district, Russia

E-mail: gshn5@yandex.ru

Abstract. Today, due to the large spread of cloud technologies, more and more individuals and businesses use their services. It is common practice in the field of cloud technologies to store information (data) from one service provider in one physical and geographical location. At the same time, we must remember that the information passes through third parties whose intentions are not always known. Therefore, valuable critical information is usually encrypted before it is sent to the cloud on the user's side. To do this, you can use both simple tools (for example, creating password-protected archives) and more advanced tools (software for creating encrypted partitions). However, in addition to normal encryption, you can use the fact that there are many unrelated repositories. Such methods include, for example, the dissection-separation method, which is discussed in this paper.

1. Introduction

It is a common practice in the information technology and communications industry to store information (data) at the same service provider, in the same physical and geographical location. Often, the service provider also provides redundancy and re-storage of data in remote locations, but this is only to ensure the performance and availability of data in the event of an accident or other failure. Information (data) is still entrusted to this single service provider, which must implement a variety of security measures to ensure the protection of information, namely its confidentiality, integrity, and availability. These measures may be preventive, investigative or corrective in nature and include, but are not limited to, physical and logical access control, data encryption at rest and during transmission, backup, monitoring, notification, logging, antivirus, firewalls, intrusion detection and prevention systems, physical and environmental protection measures, security policies, procedures and standards, data verification, security awareness programs, audit and certification plans, incident response, breach notification, risk assessment, etc.

Encryption is an important element. However, it may be poorly implemented, contain backdoors unknown to the user or data owner, and the keys themselves may be transferred to third parties without the knowledge or consent of the user or data owner. This makes it an urgent problem to develop methods that are additional or even alternative to encryption. For example, you can use the fact that there are many unrelated repositories. Such methods, in particular, include, for example, the dissection-separation method, let's consider its effectiveness and reliability in more detail.
2. The general principle of operation and the essence of the dissection-separation method

Dissection (separation) or dissection-separation of information is a cryptographic method of information protection, which consists in the fact that the array of protected data is divided into parts, each of which separately does not allow revealing the content of the protected information. These fragments can be transmitted across multiple sources, distributed over time and at the recording location on a floppy disk or any other recording device. The dissection-separation method is divided into semantic and mechanical according to its implementation [1-3].

Let us consider the following example. We take the text as the source message:

'Who_controls_the_past,’_ran_the_Party_slogan_'controls_the_future:_who_controls_the_present.controls_the_past.' (George Orwell. 1984) and let us try to evaluate the effectiveness of the method using this example. The total number of characters in the message is 110.

To split a block of plain text, we use the technique described in [2]. We cut the text into 12 parts and write the plain text in the following format (table 1).

3 rows and 4 columns are selected to split the text into 12 parts. Let sj columns be selected in the sequence {4, 1, 3, 2}, and the ri strings are in the sequence {2, 3, 1} [4].

Text characters are placed in blocks that are identified by numbers. Number k of the block F(k), where the next character of the plain text is written, is determined by the formula:

\[ k = (r_i - 1) \cdot n + s_j \]

where n is the number of columns [5]. Then the first character “W” will be recorded in the block number (ri = 2, sj = 4): k = (2 – 1) · 4 + 4 = 8; the second character “h” – block number (ri = 2, sj = 1): k = (2 – 1) · 4 + 1 = 5; the third symbol “o” in block number (ri = 2, sj = 3): k = (2 - 1) · 4 + 3 = 7; etc. (table 2).

| Table 1. Dissecting an open text |
|----------------------------------|
| * 4 1 3 2 4 1 3 2 4 1 3 2 4 1 3 2 |
| * 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 |
| 2 1 Who _ _ the r a n _ y _ s l on t r f u tu c o n t |
| 3 2 c o n t _ p a s t _ e _ o g a n o l s _ r e : _ r o l s |
| 1 3 r o l s t , ‘_ _ P a r t , _ _ c t h e _ w h o _ _ t h e |

Continuation of table 1

| Table 2. Splitting plain text (which was previously split) |
|----------------------------------------------------------|
| 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 |
| 1 8 5 7 6 8 5 7 6 8 5 7 6 8 5 7 |
| 2 12 9 11 10 12 9 11 10 12 9 11 10 12 9 |
| 3 4 1 3 2 4 1 3 2 4 1 3 2 4 |

* - sj columns are selected in the sequence {4, 1, 3, 2}
** - ri strings - in the sequence {2, 3, 1}
As a result, blocks F(k) written in number order will contain the following characters:

F1 = {o,a_hhtc_...}
F2 = {s_tc_ ena...}
F3 = {r'reohop...}
F4 = {rtp,tw_ _e...}
F5 = {hta_nuoprt...}
F6 = {e_ethu...}
F7 = {ohnsttro...}
F8 = {W_ryofc_ts...}
F9 = {ophgleoe_...}
F10 = {ts_n_ _th...}
F11 = {naes:Int...}
F12 = {c_toorss...}.

Thus, one block of plain text is replaced by twelve blocks, which together give the length of the source text. Then the resulting blocks are distributed to different storage locations, and all this is a reversible process. This implementation of the split-diversity method is called mechanical. Let us consider its persistence.

3. Evaluation of cryptographic strength

The cryptanalysis method dissection-separation will be performed for the case by the brute force method in 3 directions:

1. when an attacker knows all 12 blocks of source code that they need to collect;
2. when one knows how to implement the split-split method and 1 block of 12 source messages;
3. when one knows only 1 block.

Case 1. If the attacker has all the blocks F1-F12, he must arrange the available characters in the correct order. Then you can calculate the number of combinations of possible solutions using the iteration method:

\[
P_m = m! = 110! = 1.5882455415 \times 10^{178}
\]

where \( m \) is the total number of characters [6]. It is easy to see that the brute force method for the first case, even on such a short message, gives an excessive amount of calculation. You can reduce it a little by noting that the letters “W” and “P” are capital letters, which means that one of them should go 1st.

\[
P_m = m! = (110-1)! = 109! + 109! = 2.8877191664 \times 10^{176}.
\]

Instead of the brute force method, you can use frequency analysis if the message is long enough (in this case, it is 110 characters). It allows you to determine whether encryption was used or whether the original message was simply divided into blocks. If there was encryption, the frequency of characters in the message did not match their frequency in English, and this could also indicate the use of some other language.

Seeing that we have 17 spaces ( " _ " ) suggests that we should have 18 words. Accordingly, if we have 18 words in the message, the number of combinations will be significantly reduced. In this case, you can rely, for example, on dictionaries:
- Cambridge Advanced Learner's Dictionary - English dictionary edition for advanced users. Suitable for students with B2 Upper Intermediate - C2 High Advanced proficiency levels. Contains 140,000 words and phrases.
- Longman Collocations Dictionary and Thesaurus – English dictionary of phraseology, phrases and synonyms for users with intermediate – Advanced proficiency levels. The dictionary includes 700,000 phrases and synonyms to help improve your spoken and written British English skills.
- Oxford Advanced Learner's Dictionary - English dictionary for students and adults with B2 Upper-Intermediate – C1 Advanced proficiency. Contains 185,000 words and expressions.
- Collins English Dictionary – one of the most complete dictionaries of the English language. The dictionary covers 722,000 words and phrases.
- and other.

To do this, you can build combinations of these words and our known letters by filtering through these dictionaries. Moreover, when using statistical laws of the English language, to improve the search, you can pay attention to the most frequent bigrams and trigrams, as well as pay attention to pairs of repeated letters. In English, the most commonly repeated letters are ss, h, tt, ff, ll, mm and oo. If there are any duplicate characters in the ciphertext, we can assume that they represent one of these pairs.

If there are spaces between words in the ciphertext, we will try to identify words consisting of one, two or three letters. The only words in English that consist of a single letter are a and i. The most common two-letter words are of, to, in, it, is, be, as, at, so, we, he, by, or, on, do, if, me, my, up, an, go, no, us, am. The most commonly used three-letter words are the and and [7].

If possible, prepare a table of letter frequencies for the message we are trying to decrypt. For example, military reports tend to omit pronouns and articles, and the absence of words such as I, he, a and the will reduce the frequency of some of the most common letters [7].

If we know that we are working with a military report, we should use the letter frequency table created from other military reports.

One of the most useful skills for a cryptanalyst is the ability to recognize words or even entire phrases through their own experience or purely intuitively. Al-Khalil, one of the first Arab cryptanalysts, demonstrated his abilities when he cracked the Greek ciphertext. He suggested that the ciphertext begins with the greeting "In the name of God". Having established that these letters correspond to a certain fragment of the ciphertext, he was able to use them as a scrap and reveal the rest of the ciphertext. This was called the crib [8].

By skipping our options through these bigrams and trigrams, we can reduce the amount of iteration. There is a fairly significant probability that the original message will satisfy this option, but here, in order to better work out all the options, it is advisable to use a search for climbing to the top or an algorithm for simulating annealing. In the desired version, some of the letter combinations will be put down at the expense of bigrams and trigrams, so you can continue to deliver the missing letters based on the meaning of words, phrases, morphological rules, the approximate subject of the message, etc. If the original message is able to restore at least 5-10% of the words, and in all other words at least some letter combinations, then it is already quite possible to restore the entire message.

Case 2. In this case, when the attacker has one of the blocks (for example, F11 = {naeas:ln...}) and its number, and also knows the method by which the split-spacing method was implemented, it can be concluded that there are 12 blocks in total, since the text is separated into 3 rows and 4 columns (Table 3 and 4).

And since our 11 block has 9 characters instead of 10, we can conclude that some blocks have either 9 or 10 characters.

You can immediately put it in the table 4 after the dash in block 11, there are also dashes in subsequent blocks.

3 rows and 4 columns are selected to split the text into 12 parts. Let sj columns be selected in the sequence {4, 1, 3, 2}, and the ri strings are in the sequence {2, 3, 1}. 
Then the number \( k \) of the block \( F(k) \), where the next character of the plain text is written, is determined by the formula:

\[
k = (r_i-1) \cdot n + s_j
\]

where \( n \) is the number of columns.

In principle, we can restore a table with blocks (Table 5).

All letters of the alphabet are in upper and lower case \( 52 + 1 \) (space) + 10 numbers (0-9) + 12 punctuation marks ( –, . ! ? : ; “ ” ( ) ’ ... ). A total of 75 is possible options per 1 cell. There are 114 cells in total (minus the known 6 empty cells in table 4).

| Table 3. Dissecting an open text |
|----------------------------------|
| 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 |
| 1 |
| 2 \[ \begin{array}{c}
  \text{n} \\
  \text{a} \\
  \text{e} \\
  \text{a} \\
  \text{s} \\
  \text{;} \\
  \text{l}
\end{array} \] |
| 3 |

Continuation of table 3

| Table 4. Splitting plain text (which was previously split) |
|----------------------------------------------------------|
| 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 |
| 1 |
| 2 \[ \begin{array}{c}
  \text{n} \\
  \text{t} \\
  \text{-} \\
  \text{-}
\end{array} \] |
| 3 \[ \begin{array}{c}
  \text{-} \\
  \text{--} \\
  \text{-} \\
  \text{-} \\
\end{array} \] |

Let's use the placement formula with repetitions.

\[
\text{A}_n^k = n^k = 75^{114} = 5,7145761808 \times 10^{213}, k \text{ – total cells where elements with repetitions are placed, } n \text{ - all types of placed elements [9].}
\]

It is necessary to exclude all variants with repeated letters and punctuation marks 2 or more times in a row, this does not happen in English with the exception of the double spelling of the most frequently repeated letters: ss, ee, tt, ff, 11, mm and oo.

It seems that this will significantly reduce the search. \( n^k_1(65) = 65^{114} = 4,7002683899 \times 10^{206} \), this is a combination of punctuation marks and letters that should not be repeated or can be repeated, but with rare exceptions compared to the total number of variants. It can be assumed that letters and punctuation marks can be repeated once or twice or more. Then you can search for the original message in the same way as in the previous version, starting with bigrams and trigrams.
Case 3. The attacker has one of the blocks, for example \( F_{11} = \{ \text{naeas:ln} \ldots \} \). In this case, a favorable outcome is very unlikely. The first possible option, if there are only 2 blocks of equal length, then we get:

\[
\frac{\text{m}}{\text{n}} = \frac{\frac{\text{m}}{\text{n} + \text{m} - 1}}{\text{m}! (\text{n} - 1)!}
\]

\( \text{n} \) – all known elements, \( \text{m} \) - variants of the arrangement [10].

\[
P_1 = (\text{n}! \cdot \frac{\text{m}}{\text{n}})^2 = (9! \cdot \frac{8!}{7!5!8!}) = 1,4313158201 \times 10^{16}
\]

\[
P_2 = 1,4313158201 \times 10^{16} \cdot \frac{8!}{7!5!8!} = 5,6455714754 \times 10^{26}
\]

etc.

Then you need to take into account that there may be more than 2 blocks. And in 2, 3 and further blocks, the number of characters can be more or less, which will constantly increase the number of options many times. Exponentially.

Moreover, the number of meaningful words that can mean our original search message will constantly increase, also exponentially. For successful implementation of the search, you will still need additional information, such as the approximate meaning of the message or what preceded this message (user correspondence), etc.

Table 5. Splitting plain text (which was previously split)

| 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 8 | 5 | 7 | 6 | 8 | 5 | 7 | 6 | 8 | 5 | 7 |
| 2 | 12 | 9 | 11 | 10 | 12 | 9 | 11 | 10 | 12 | 9 | 11 |
| 3 | 4 | 1 | 3 | 2 | 4 | 1 | 3 | 2 | 4 | 1 | 3 |

Continuation of table 5

| 1 | 8 | 5 | 7 | 6 | 8 | 5 | 7 | 6 | 8 | 5 | 7 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 12 | 9 | 11 | 10 | 12 | 9 | 11 | 10 | 12 | 9 | 11 |
| 3 | 4 | 1 | 3 | 2 | 4 | 1 | 3 | 2 | 4 | 1 | 3 |

Continuation of table 5

| 1 | 8 | 5 | 7 | 6 | 8 | 5 | 7/- | 6/- |
|---|---|---|---|---|---|---|---|---|
| 2 | 12 | 9 | 11 | 10 | 12/- | 9/- | - | - |
| 3 | 4 | 1 | 3 | 2 | - | - | - | - |

4. Discussion

In addition to this split-spread method, there are such "classical" encryption methods as substitution, permutation, and XOR cipher [11-13]. They are inherently linear in the sense that the length of the encrypted message is equal to the length of the source text. Non-linear transformation of the substitution type is possible instead of the original characters (or whole words, phrases, sentences) of pre-selected combinations of characters of different lengths [12-14]. Also recently, the so-called computer steganography (from the Greek words steganos - secret, secret and graphy - record), which is hiding a message or file in another message or file, has become popular [15, 16]. For example, you can hide an encrypted audio or video file in a large information or image file.

But in this article, first of all, we consider the possibility of using different storage locations in the "clouds" or any other media. The only difference is that the above numerical methods allow you to store the modified source file in one storage location, and backup copies in another. And if the attacker can get this file and hack it, then the goal of the crypto attack will be achieved. The dissection-separation method considered in this article allows you to split the source file into several parts and store each in different storage locations. if an attacker can get one part and successfully hack it, then they will still not be able to achieve the goals of the crypto attack. This will only be possible under favorable circumstances, for example, if it can extract all parts of the source file or most of them.
From the above, we can conclude that this method can be effectively hacked under certain conditions.

If we turn to the well-known cryptographer Auguste Kerkhoff, who is the author of the fundamental work "Military cryptography" ("La Cryptographie Militaire"), in which he formulated General requirements for the cryptosystem, and also approved cryptanalysis as the only true way to test ciphers [17].

Six general criteria are for a cryptographic system.
1. The system must be physically, if not mathematically, undetectable.
2. It should not be necessary to keep the system secret; getting the system into the hands of the enemy should not cause inconvenience ("the Kerkhoff Principle").
3. Key storage and transfer should be possible without the use of paper records; correspondents should be able to change the key at their discretion.
4. The system must be suitable for Telegraph communication.
5. The system should be easily portable, and working with it should not require the participation of several people at the same time.
6. Finally, the system is required, given the possible circumstances of its application, to be easy to use, not require significant mental stress or compliance with a large number of rules [18, 19].

For points 1 and 2, this split-split method cannot be considered reliable. This method can be considered according to these six criteria, despite the fact that it does not belong to the cryptosystem, but only a method. Because it can be included in the cryptosystem itself, which can increase the risks of a successful cryptographic attack, in view of its vulnerability, and increase the cost of improving its reliability.

If we turn to the doctrine of the famous French cryptologist Antoine Rossignol, according to which the strength of the military field cipher must be such as to ensure the secrecy of the cipher message until it is received by an army unit and the order is executed. A diplomatic cipher must be such that it takes several tens or even hundreds of years to reveal it, since in diplomacy, an encrypted document must often remain secret for a very long time after it is encrypted [20]. Today, this doctrine can be paraphrased as follows: a prerequisite for today's ciphers is non-disclosure not only in diplomacy, but also in military Affairs and in other areas of human and managerial activity, in particular when working with critical information infrastructures.

5. Conclusion
From the work done, we can conclude that this cryptographic method of protecting information in external data warehouses is very interesting and meets the "spirit of the time", but the dissection-separation method itself is significantly vulnerable. This is confirmed by the identified two directions in this article: knowledge of the method of implementation of this method, knowledge of all the threads into which the split source message was distributed. The topic and direction need to be further explored.

6. References
[1] Bezbokov A A, Yakovlev A V and Simkin V N 2006 Methods and means of computer information protection: a Training manual, available at: https://tstu.ru/book/elibri/pdf/2006/shamkin2.pdf
[2] 2010 Cryptogram. A practical guide to manual encryption. - 4th ed. (Saratov: Novy Veter publishing house)
[3] Kondratenko S V and Novikov Yu V 2005 Basics of local networks. National research nuclear
(University "MEPhI")
[4] Partyka T L and Popov I I 2002 Information security. Training a textbook for students
of institutions of secondary vocational education (Moscow; FORUM: INFRA-M)
[5] Lygin E A 2010 L 88, Secret Writing. Practical guide to manual encryption (Saratov: Novy
Veter publishing house)
[6] Vilenkin N Ya 1969 Combinatorics (Moscow: Nauka)
[7] Cryptographic methods and data protection tools, available at:
http://develab.narod.ru/crypt/crypt04.htm
[8] Other encryption methods, available at:
http://www.univer.omsk.su/omsk/Edu/infpro/1/kodir/drmetsh.html
[9] Ryser H J 1963 Combinatorial Mathematics, The Carus Mathematical Monographs (#14) (The
Mathematical Association of America)
[10] Riordan J 1958 An Introduction to Combinatorial Analysis (New York, Wiley)
[11] Ravi Gupta and Rahul Singh 2015 SSCC - An Improved Substitution Method for Data
Encryption Using DNA Sequence and CDMB. Communications in Computer and Information
Science, pp. 197-206 DOI http://dx.doi.org/10.1007/978-3-319-22915-7_19
[12] khlas Abbas Albahrani, Tayseer Karam Alshekly 2017 New Chaotic Substation and
Permutation Method for Image Encryption International Journal of Applied Information Systems
12(4) 33-39 DOI http://dx.doi.org/10.5120/ijais2017451698
[13] Fei Huo, Guang Gong 2015 XOR Encryption Versus Phase Encryption, an In-Depth Analysis
IEEE Transactions on Electromagnetic Compatibility, 57(4) 903-911 DOI
http://dx.doi.org/10.1109/temc.2015.2390229
[14] Muhammed Jassem Al-Muhammed 2018 Light but Effective Encryption Technique based on
Dynamic Substitution and Effective Masking International Journal of Advanced Computer Science
and Applications 9(9) DOI https://dx.doi.org/10.14569/IJACSA.2018.090909
[15] Sefa Tuncer 2016 Information Encryption and Hiding into an Image by Steganography Methods
to Improve Data Security Journal of new results in science 5(12) 170-177
[16] Gotfried C Prasetyadi, Achmad Benny Mutiara, Rina Refianti 2017 File encryption and hiding
application based on advanced encryption standard (AES) and append insertion steganography method
Second International Conference on Informatics and Computing (ICIC), Nov 1, 2017. DOI
http://dx.doi.org/10.1109/iac.2017.8280584
[17] Kahn D 1996 The Codebreakers.
[18] Kerckhoffs’ principles, Available at:
https://www.petitcolas.net/kerckhoffs/la_cryptographie_militaire_i.htm#desiderata
[19] Kerckhoffs A 1883 La cryptographie militaire. Journal des sciences militaires
[20] Perrault Ch 1697 Les Hommes illustres (Paris: Dezaillier)