Cilt 11, Sayı 4, Sayfa 135-141, 2022 Volume 11, Issue 4, Page 135-141, 2022

Araştırma Makalesi https://doi.org/10.46810/tdfd.1208075 Research Article

Türk Doğa ve Fen Dergisi
Turkish Journal of Nature and Science
www.dergipark.gov.tr/tdfd

Behavioral Steganography in Social Networks

Muharrem Tuncay GENÇOĞLU*

[1] Fırat University, Vocational School of Technical Sciences, Elazığ, Türkiye
Muharrem Tuncay GENÇOĞLU ORCID No: 0000-0002-8784-9634

**Corresponding author: mt.gencoglu@firat.edu.tr*

(Received: 21.11.2022, Accepted: 13.12.2022, Online Publication: 28.12.2022)

Keywords
Steganography, behavior Steganography, hide the existence of the information, Knapsack algorithm, Social network

Abstract: Recently, using human behavior to hide the existence of information has been at the center of steganography research. In this study, a behavioral steganography algorithm using CMI (Coded Signal Inversion) coding is proposed to minimize the high bit error rate that occurs when transmitting a large number of continuous and identical confidential information in the knapsack algorithm, which is used to improve information transmission efficiency and flexibility of transmission mode in social networks. In the proposed algorithm; Data redundancy is reduced by reducing the number of mutual friends of the sender and each receiver. Then, the proposed algorithm was applied and the results were analyzed. Experimental analysis shows that this scheme improves the practical value of behavioral steganography in social networks and has high security.

Sosyal Ağlarda Davranışsal Steganografi

Anahtar Kelimeler
Steganografi, davranış Steganografisi, bilginin varlığını gizleme, Sırt çantası algoritması, Sosyal ağ

Öz: Son zamanlarda bilginin varlığını gizlemek için insan davranışlarının kullanılması steganografi araştırmalarının merkezinde yer almaktadır. Bu çalışmada, sosyal ağlarda iletim modunun verimliliği ve esnekliğini iyileştirmede kullanılan knapsack algoritmasında çok sayıda sürekli ve özdeş gizli bilgi iletilirken oluşan yüksek bit hata oranını en aza indirmek için CMI (Coded Signal Inversion) kodlaması kullanan bir davranışsal steganografi algoritması önerilmiştir. Önerilen algoritmada; Göndericinin ve her alıcının ortak arkadaşlarının sayısı azaltılarak veri fazlalığı azaltıımıştır. Daha sonra önerilen algoritma uygulanmış ve sonuçlar analiz edilmiştir. Deneysel analiz, bu şemanın sosyal ağlarda davranışsal steganografinin pratik değerini geliştirdiğini ve yüksek güvenliğe sahip olduğunu göstermektedir.

1. INTRODUCTION

Steganography [2,6], known as hiding the existence of data, can be defined as the technology of using information embedding algorithms to hide confidential information on non-secret networks. Continuous improvements in steganalysis techniques reduce the reliability of traditional steganography day by day. The decrease in the reliability of steganography emerges as a serious problem in this field. To eliminate this problem, steganography methods applied against steganalysis have emerged. Recently, using human behavior to hide the existence of information has been at the center of steganography research. Social networks have become an integral part of life today. Especially services such as Facebook, Instagram, Google, and LinkedIn have had an important place in interaction and communication in recent years.

The most important communication features in social networks are time and interaction. Based on these features; Pantic and Husain proposed a social network behavior steganography using the length of Twitter information to convey confidential information [11]. Li et al. used selected social networks and online accounts to hide the existence of information [9] Zhang suggested a similar behavior technique using behaviors in social media [16]. However, this method is very difficult to implement due to the high bit error rate in the transmission of confidential information containing many consecutive identical bits. Zhao et al. proposed an FPGA-based CMI (Coded Mark Inversion) design [17]). Subramanian et al.

proposed a channel-based binary coding technique for secure data transmission in wireless networks [12]. This method is not very useful as it proposes a distributed methodology with a high compression ratio. Çıtlak et al. analyzed the existing methods for detecting spam accounts in the Twitter network and compared the strengths and weaknesses of the methods for distinguishing real users and fake users [1]. Hu et al. proposed a behavioral correlation-based stenographic method for social networks [6]. Kantartopoulos et al. analyzed hostile attacks based on AdaBoost on fake Twitter accounts using machine learning and suggested the use of K-NN for defense [7]. Li et al. proposed a data hiding technique that transforms a secret message directly into a hologram-based fingerprint image obtained from the secret message [8].

The motivation for this study is to reduce the bit error rate mentioned above. For this, a 0-1 knapsack algorithm based on the probabilistic solution finding algorithm [5] proposed by Hu for solving the 0-1 knapsack problem will be proposed.

Firstly, the 0-1 knapsack problem will be introduced in this context. Then, behavioral steganography algorithm flow based on the 0-1 knapsack algorithm will be given. Then, the experimental results of the feasibility analysis will be given. Finally, the findings of the study will be evaluated.

2. 0-1 KNAPSACK PROBLEM

Wang at all. the genetic algorithm is used to solve the 0-1 knapsack problem and the principles and implementation process of the two methods are analyzed [13,14]. Han and Li applied a chaotic transformation with chaotic map image encryption [3,4]. Hu et al. proposed behavioral steganography based on the 0-1 knapsack algorithm. They used CMI coding to solve the high bit error rate when transmitting a large number of continuous and identical confidential information [6].

The Knapsack problem simply aims to fit the most items in a bag. In the 0-1 knapsack problem; All items are either bought or left. It is not possible to take part in the item to be purchased. Therefore, if we indicate with X_i whether an item is bought or not, the problem can be modeled as follows:

$$\sum_{i=1}^{n} p_i x_i$$

to be as large as possible and $x_i \in \{0,1\}$

$$\sum_{i=1}^{n} a_i x_i \leq c_i.$$

In this model, the value of X_i can be 0 or 1. If it is 0, it is not taken from the i element, and when it is 1, the whole element i is taken.

2.1. Coded Mark Inversion (CMI)

"CMI encoding doubles the data rate. A zero is sent as a low to high [01] transition, while a one is sent as either a one (1) or zero (0) depending on the previous state. If it was low the one is sent as a one (1)" [18].

2.2. Arithmetic Coding (AC)

"The basic idea of arithmetic coding is to use a range of numbers between 0 and 1 to represent each possible series of n messages" [10].

3. Behavior Steganography Based on the Proposed Knapsack Algorithm

The detailed process of the proposed behavioral steganography method based on the 0-1 knapsack algorithm between sender and receiver is as follows:

3.1. Sender

To be represented by the Sender S;

Step 1.
$M = [m_1, m_2, \dots, m_l]^T$ binary secret
N = Number of friends of the sender
H_i = similarity matrix that records friend's likes on the sender's posts
In the H matrix, liking is recorded as 1, and dislike as 0.

$$\begin{bmatrix} v(1,1) & v(1,2) & \cdots & v(1,l) \\ \vdots & & \ddots & \vdots \\ v(n,1) & v(n,2) & \cdots & v(n,l) \end{bmatrix}_{nxl} \quad (1)$$

Step 2.
i: Friend, j: Like status;
The sender sends a secret M message R_i ($i = 1,2,\dots,r_1$) to the r_1 receivers. CMI coding is done on M to get rid of consecutive similar bits in M and to obtain the N column [17].

Step 3.
By multiplying H and N, the D post sequence is obtained. G matrix is obtained by encoding the D post matrix with inverse arithmetic coding. Thus, decimal numbers between 0 and 1 with probability p are represented as information strings [12]. The probability P is derived from the H matrix. 1 ratio in rows of H; A probability of 1 gives a probability of 0, a probability of 0. Multiply D by a coefficient b, such as 10^{-1} or 10^{-2}, to get all probabilistic decimals.

$$D = \begin{bmatrix} v(1,1) & v(1,2) & \cdots & v(1,l) \\ \vdots & & \ddots & \vdots \\ v(n,1) & v(n,2) & \cdots & v(n,l) \end{bmatrix}_{nxl} X \begin{bmatrix} 1 \\ 2 \\ \vdots \\ l \end{bmatrix}_{lx1} = \begin{bmatrix} 1 & 2 & \dots & n \end{bmatrix}_{nx1} \quad (2)$$

Step 4.

The sender marks the tracking movements of n friends according to the G matrix. That is if the element in G is 1, he likes it, if it is 0, he does not like it.

3.2. Receiver

Recipients R_i (i=1,2,…,r_1) and r_1 are the number of recipients, and non-recipients are N_j (j=1,2,..,r_2) and r_2 are the number of non-recipients;
R_i and N_j make possible contacts. well
$K=r_1+r_2$ (3)
All persons are expressed with p_λ $(\lambda = 1,2,\dots,k)$.

Step 1.
R_i's are selected in p_λ. R_i; It reconstructs the G matrix according to the likes, creating the G_i matrix corresponding to the mutual friends of S and R_i. R_i scans the G_i for rows, removes the same rows, and represents the final matrix with A.

Step 2.
The matrix B is obtained by coding the arithmetic according to the probability P of A. The matrix B is multiplied by b^{-1}. A is a submatrix of G. Since each row of G corresponds to H, R_i observes H to form the matrix corresponding to A and denotes it by C. So C becomes the submatrix of H. N columns are obtained by reconstructing the data, and *M* message is decoded by performing CMI decoding on N. If the receiver's status is represented by a y value, the sender can check the receiver's status based on that y value.
w_λ , is to show the weight of p_λ;
if p_λ is the receiver w_λ>0
If p_λ is not a receiver, let w_λ=y.

$$\sum_{\lambda=1}^{r_1+r_2} w_\lambda = y \quad (4)$$

y will reflect the state of p_λ.

The sender sets a threshold value of $a > y$ and changes the value of a to check the status of p_λ. The relationship between a and y is denoted by

$$\operatorname{argmin}(a-y) = y \quad (5)$$

When $x_\lambda = 1 p_\lambda$ is a receiver,

When $x_\lambda = 0$, p_λ is not a receiver. This relationship is mathematically

$$\sum_{\lambda=1}^{r_1+r_2} w_\lambda x_\lambda \le a \quad (6)$$

is expressed by the correlate. That is, the sender selects a threshold to check the p_λ state and uses the minimum difference between a and y for the optimal state of p_λ. Each of p_λ weights of w_λ and a value of v_λ.
The receiver chooses 〖$r_1 p_\lambda$ randomly among $k p_\lambda$ and values for $v_{k\varphi}(\varphi = 1,2,\dots,r_1)$

$$\sum_{\varphi=1}^{r_1} v_{k\varphi} = v_{[k][c]} \quad (7)$$

checks whether the formula is met. Here $v_{[k][c]}$ is the maximum value of the knapsack problem. If equation (7) is satisfied, all $r_1 p_\lambda$ are receivers, otherwise, the 0-1 knapsack assignment protocol is repeated. The 0-1 knapsack assignment protocol is shown in figure 1.

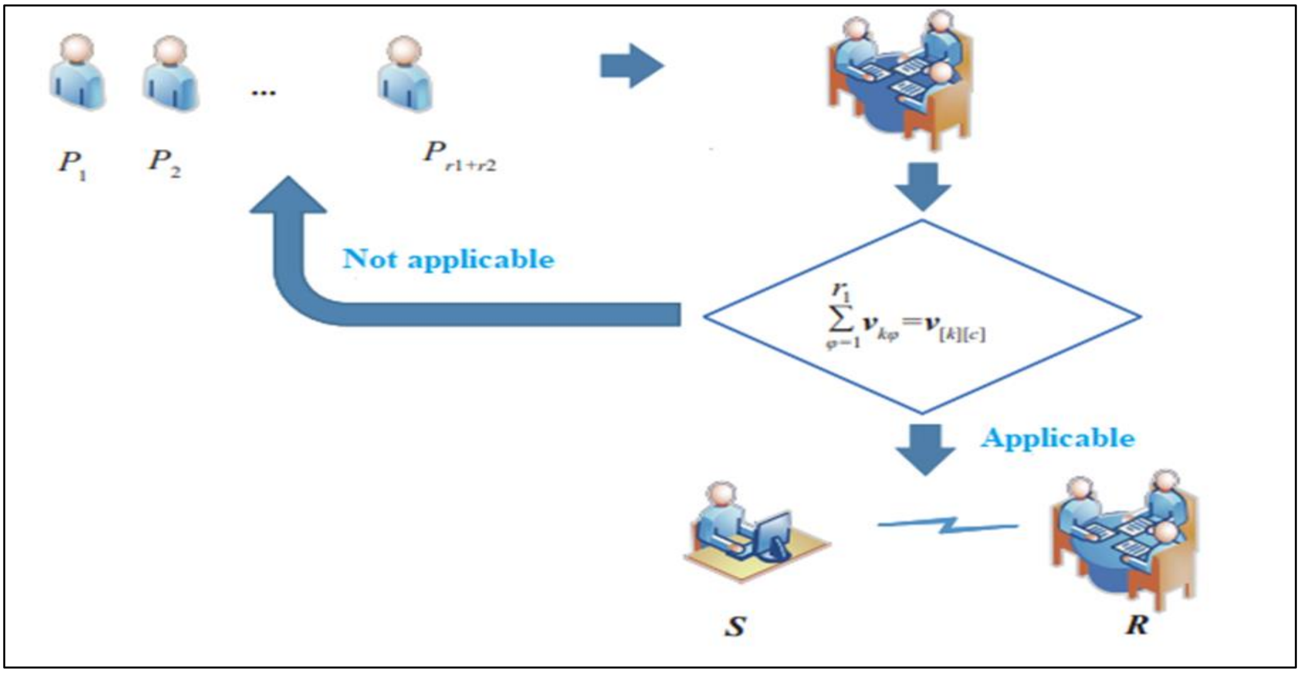


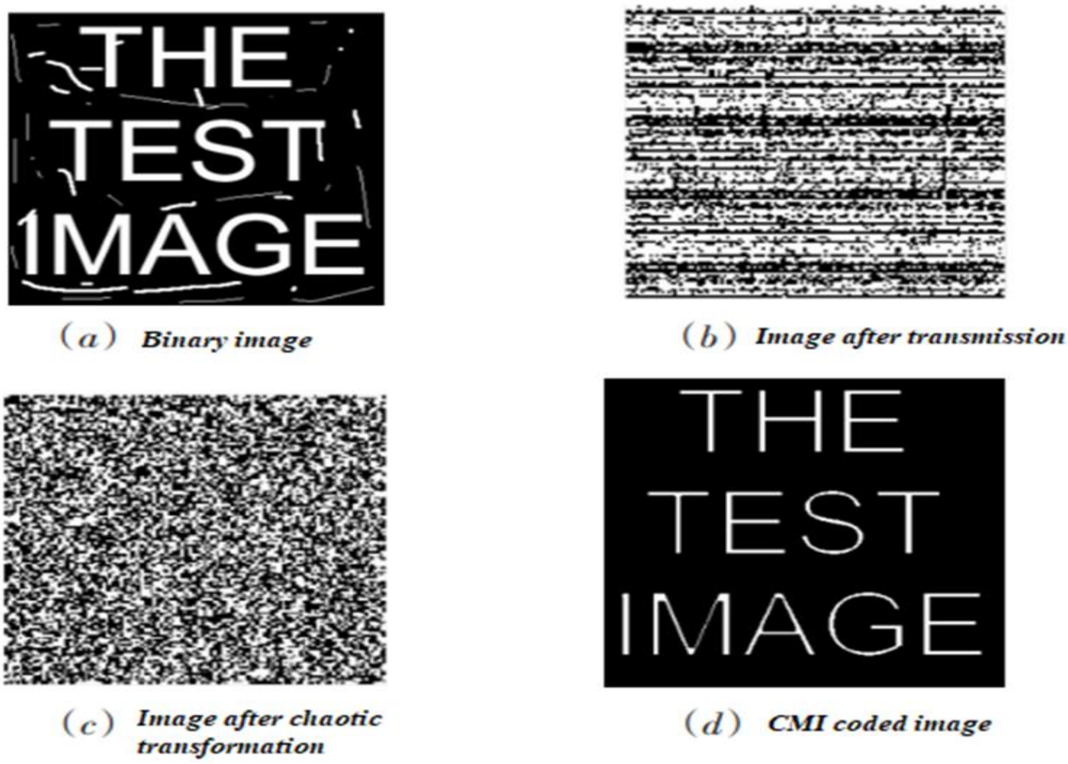
Figure 1. Knapsack personnel assignment protocol

2.3. Algorithm Flow

The detailed process of behavioral steganography based on the proposed 0-1 knapsack algorithm is as follows:

Behavioral steganography algorithm
Input: Secret information *M*, like matrix H, like probability P, knapsack capacity C, value information $V=[v_1,v_2,\ldots,v_k]$, weight information $W=[w_1,w_2,\ldots,w_k]$, $V_{[0][\ ]}=0$, $V_{[\ ][0\ ]}=0$
Output: Secret message *M*
1. Begin CMI encoding on *M* to obtain N;
2. D ← H*N
3. calculate the inverse arithmetic coding of b.D according to P to obtain G;
4. randomly select k sub-matrices of G as the common friends of the sender and k receivers;
5. for i ←1 to k
6. do for j ←1 To c
7. v[i][j] ← v[i - 1][j] ;
8. If (j ≥ wi)
9. v[i][j] ← Max (v[i][j], v[i - 1][j - w_i]+ v_i) ;
10. return v [k][c] ;
11. randomly select r_1 elements in $V=[v_1,v_2,\ldots,v_k]$ (7) is satisfied
12. connect the matrix G_i in turn, and remove the same row of G_i to get the matrix A;
13. find the C corresponding to A in H;
14. calculate the arithmetic code of A according to P, and multiply the result by b^{-1} to get B;
15. reconstruct the solution N according to the data;
16. decode N with CMI to get *M*
17. end

The sender corresponds to lines 1-4 in the pseudocode, the receiver corresponds to lines 5-16 in the ps4.



(*a*) ***Binary image*** (*b*) ***Image after transmission*** (*c*) ***Image after chaotic transformation*** (*d*) ***CMI coded image***

Figure 2. Feasibility Analysis of Binary Image Transfer

3. EXPERIMENTAL RESULTS

In this section, a user on LinkedIn is selected as a sender, and his friends who meet the conditions are selected as recipients. Tests and analyzes were carried out on these.

3.1. Feasibility Analysis

The 8 friends of S are designated as possible recipients, the bit number of *M* is 16 and the bit number of CMI is 8. First, in a 128x128 binary image transmission examination, 16 pixels are selected as secret information, and 6 bits with at least 8 common friends with S are selected as possible recipients. The sender selects 4 receivers and 2 non-receivers according to the 0-1 knapsack protocol. It splits the 256x256 grayscale image into 8 binary images and performs cross processing. The receiver synthesizes the 8 restored binary images into a grayscale image for bit error rate analysis.

3.2. Feasibility Analysis of Binary Image Transfer

The binary image to be transmitted is given in figure 2(a). The image after the transmission is shown in figure 2(b). Here, the bit error rate is calculated as 48.78%, which is quite high. In this binary image, the arithmetic coding has a greater error rate as there are many consecutive identical bits.

The image as a result of the chaotic transformation is given in figure 2(c). The bit error rate in this image was 51.23 percent. This rate is quite high. The CMI-coded image is shown in figure 2(d). The bit error rate in this image is calculated as 0.03%. Therefore, it is clear that the image obtained with the added CMI code gives better results since it has a low bit error rate.

3.3. Feasibility Analysis of Transferring Grayscale Image

The Lena grayscale image is shown in figure 3(a). The 8 bitmaps separated from the grayscale image are shown in Figure 3(b) and Figure 3(i), respectively. The image separated from the grayscale image after the receiver correction is seen in figure 3(j). The total bit error rate of 8 bitmaps was calculated as 0.46% and the bit error rate of grayscale images was calculated as 2.99%. These experimental results revealed that the proposed method has a low bit error rate and high applicability.

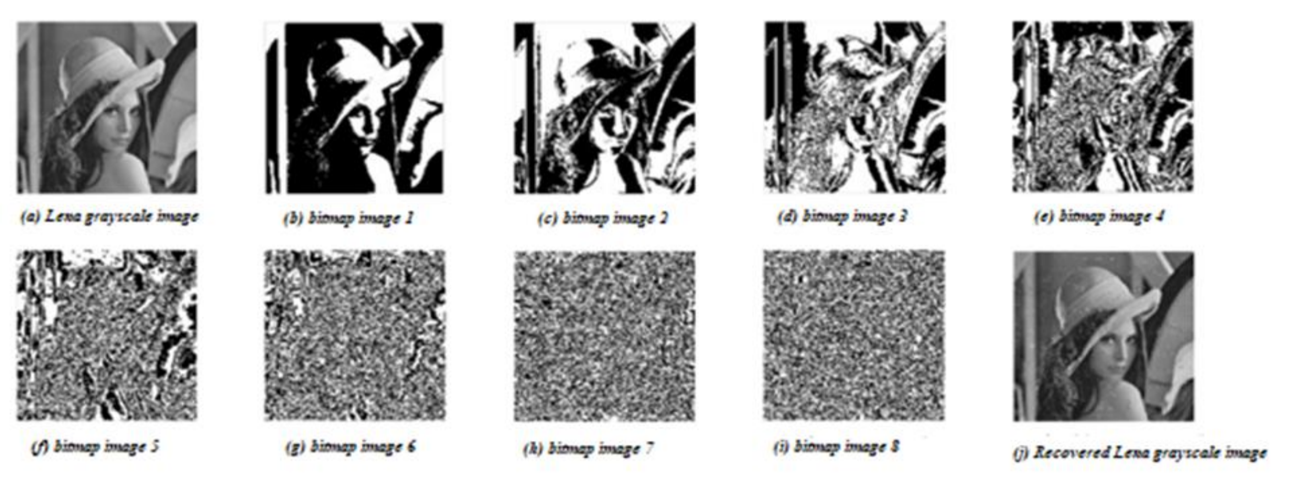


Figure 3. Feasibility Analysis of Grayscale Image Transfer

4. SECURITY ANALYSIS

4.1. Security Analysis Under Brute-Force Attacks

Brute-Force attacks [15] are successful code-breaking attacks that usually target limited capacity password dictionaries and encryption schemes. Although behavioral steganography on social networks has a limited capacity password dictionary, they are resistant to Brute-Force attacks. Therefore, the proposed behavioral steganography is theoretically absolutely safe against Brute Force attacks. The steganalysis algorithm [6] used to detect anomalous similar behaviors using behavioral correlation ranks the actions in the social network according to the increasing number of likes. The difference between d_1 making up less than one-third of the likes, d_2 making up more than two-thirds of the likes, and d_1^l and d_2^l the number of likes made by the two suspects, respectively, and the difference is; It is denoted by $d = \frac{d_1^l}{d_1} - \frac{d_2^l}{d_2}$. If $d < 0.25$, similar behavior is abnormal; otherwise, it is normal. Brute-Force attack and steganalysis methods were used to attack and break the proposed behavioral steganography under different test sets. The statistics of the data set used are shown in Table 1.

Table 1. Sample Statistics of Data Set

Sample Number in the Data Set	Number of Common Friends
50	10
60	12
70	13
80	15
90	17

It is known in the literature that the number of mutual friends between the sender and the receiver increases with the increase in the data set, which will lead to an increase in the total brute force cracking attack [16]. With this data set, in the brute-force attack performed under the 0-1

knapsack algorithm; It was seen that the total capacity of the brute-force dictionary did not change because the mutual friends of the sender and the receiver did not change.

t; To observe the improvement of the security of the 0-1 knapsack algorithm, provided that it shows the capacity of the brute force dictionary, a security analysis was made with different data and the results of this analysis are given in Table 2.

Table 2. Security Analysis with Different Data Sets

Sample Number in the Data Set	h	T
50	3	6545
60	4	124830
70	5	1110785
80	6	4947125
90	7	23853520
100	8	69606485

As a result, with the increase of h, it is seen that the capacity of the brute force dictionary has increased significantly and its security has increased accordingly.

4.2. Feasibility Analysis according to Receivers

For the receiver to correctly extract the confidential information M_{lX1}, the combination of transmission matrices received by other receivers must be less than $2l$ rows.

R_i ($i=1,2,\ldots,r_1$) receiver and S sender's mutual friends number n_i ($i=1,2,\ldots,r_1$);

h; It is the arithmetic mean of the number of mutual friends between R_i and S.

$$h = \sum_{i=1}^{r_1} \frac{n_i}{r_1} \quad (8)$$

Taking the data set samples given in Table 1, 4 confidential messages were sent by the sender to 5 friends, provided that the receiver was 3 mutual friends. This process was repeated 1000 times and the results were tested. The results of some of these tests are given in Table 3.

Table 3. Analysis of Receiver ($l = 5, r_1 = 4$)

S	h	h.S	Sonuç ($> 2l$)
2	3	6	$<$
	4	8	$<$
	5	10	$=$
	6	12	$>$
	7	14	$>$
	8	16	$>$
3	3	9	$<$
	4	12	$>$
	5	15	$>$
	6	18	$>$
	7	21	$>$
	8	24	$>$

In this test, it was observed whether the key point about whether the confidential information was transmitted successfully was h.S>2l, not the number of receivers. Therefore, in practice, h must be increased to achieve a high rate of successful transmission of confidential information.

4.3. Security Analysis under Impersonation Attacks

The resistance of the proposed method against impersonation attacks [7] has been tested.
In an impersonation attack; The non-recipient receives the recipient's information and acts as the recipient. Since the process of selecting r1 potential buyers is random in the protocol, there is a probability that the fraudster has successfully impersonated a certain identity.
However, for the proposed protocol, even if the attacker is successful, since he cannot be involved in the communication process, he cannot receive and destroy confidential information. Because this attack has nothing to do with the sample size of the dataset, the attacker hijacks the send matrix. This messes up the matrix and can destroy confidential information.

The analysis process is as follows;

Since the impersonation attack has nothing to do with the sample size of the dataset, the effect of the sample size of the dataset is not taken into account.

r_3 : Number of attackers among non-recipients
p_1 : the probability of success of r1 buyers
p_2 : the probability of the attacker being destroyed
p_3 : the probability of detection of the attacker and interruption of communication

Each event is tested 1000 times, and each time the recipients and non-recipients are randomly selected.

There are three conditions in each experiment:

1. Buyer succeeds
2. The attacker is destroyed
3. During the execution of the protocol, if there are two or more potential recipients with the same knapsack information, the attacker is detected and communication is interrupted.

If one of these three conditions is not met, the protocol is repeated. When the number of attackers is two, the attacker can have two situations;

1. Different receiver
2. Same recipient

These two situations are shown in Table 4 on the right and left of the cells, respectively.

Table 4. Security Testing under Impersonation Attacks

r1	r2	r3	P1 (%)	P2 (%)	P3 (%)
3	2	1	23.4	21.1	63.7
3	2	2	2.5 / 7.0	6.8 / 22.3	25.3 / 41.2
4	2	1	17.5	17.8	74.7
4	2	2	1.2 / 4.9	2.8 / 16.0	21.8 / 57.7

These tests show that the resistance of the proposed protocol against impersonation attacks is quite weak.

5. RESULT

In this study, the social steganography model of the social network 0-1 knapsack algorithm is proposed. While the proposed protocol has good performance against brute force attacks, its security against impersonation attacks is quite low. The capacity to transmit confidential information should be improved.

Acknowledgments

Muharrem Tuncay Gencoglu has been supported by TUBİTAK (121E323) also a member of the working group COST Action CA18232 Mathematical models for interacting dynamics on networks, Belgium.

REFERENCES

[1] Çıtlak O, Dörtler M. and Doğru, İ. A. A survey on detecting spam accounts on Twitter network. Soc.Netw. Anal. Min. 2019, 9:35.
[2] Dutta H, Das R K, Nandi S, An overview of digital audio steganography. IETE Technical Review, 2020, 37(6): 632 - 650.
[3] Evsutin O, Melman S, Meshcheryakov R, V. Digital steganography and watermarking for digital images: a review of current research directions. IEEE Access, 2020, 8: 166589 - 166611.
[4] Han X, Li G. Dynamic cat transformation and chaotic mapping image encryption algorithm.

Computer Engineering and Design, 2020, 41(08): 2381 - 2387.
[5] Hu F. A probabilistic solution discovery algorithm for solving 0-1 knapsack problem. International Journal of Parallel, Emergent, and Distributed Systems, 2018, 33(6): 618 - 626.
[6] Hu Y, Wang Z, Zhang X. Steganography in social networks based on behavioral correlation. IETE Technical Review, 2020, 38(1): 93 - 99.
[7] Kantartopoulos P, Pitropakis N, Mylonas, A. Exploring adversarial attacks and defenses for fake Twitter account detection. Technologies, 2020, 8 (4): 64.
[8] Li S, Zhang X. Towards construction based data hiding: from secrets to fingerprint images. IEEE Transactions on Image Processing, 2019, 28(3): 1482 - 1497.
[9] Li S, H, Wang Z, Lost in the digital wild: hiding information in digital activities. Proceedings of the 2nd International Workshop on Multimedia Privacy and Security. Toronto, Canada: Associate for Computer Mair Coolinghinery, 2018: 27 - 37.
[10] Mesut A, Veri Sikiştirmada Yeni Yöntemler (Doktora Tezi), Trakya Üniversitesi Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, 2006. http://dspace.trakya.edu.tr/xmlui/bitstream/handle/trakya/590/0042520.pdf?sequence=1, Erişim 15.04.2022
[11] Pantic N, Husain M. Covert botnet command and control using Twitter. Proceedings of the 31st Annual Computer Security Applications Conference. Los Angeles: Associate for Computer Mair coolinghinery, 2015: 171 - 180.
[12] Subramanian B, Yesudhas H, Enoch G. Channel-based encrypted binary arithmetic coding in a wireless sensor network. Ingénierie des Systèmes d'Information, 2020, 25(2): 199 - 206.
[13] Wang Z, Zhang X, Yin Z. Joint cover-selection and payload-allocation by stenographic distortion optimization. IEEE Signal Process Lett, 2018, 25(10): 1530 -1534.
[14] Yan W, Min W, Jia L, Xiang X. Algorithmic decision analysis of 0-1 knapsack problem. Computer Knowledge and Technology, 2020, 16(04): 259 - 264.
[15] Zhang W, Qin Z, Feng Z, Liu J, Liu W, Tang X. Big data analysis for detection of web brute-force attack. Journal of Shenzhen University Science and Engineering), 2020, 37(S1): 44 - 49.
[16] Zhang X. Behavior Steganography in Social Network. Taiwan, China: Springer International Publishing, 2017: 21 - 23.
[17] Zhao X, Cheng Y, Zuo L, Fang Y. Design of CMI CODEC based on FPGA. Modern Information Technology, 2020, 4(19): 35 - 37.
[18] http://www.interfacebus.com/CMI_Encoding_Definition.html, Erişim: 15.04.2022