Security-preserving social data sharing methods in modern social big knowledge systems

Xuan Chen

Security-preserving Social Data Sharing Methods in Modern Social Big Knowledge Systems

Xuan CHEN[a]

[a]*Sorbonne University, Paris, 75005, France*

Abstract

In recent decades, the development of social computing systems has realized the efficient information exchange between large groups of people. Nowadays, social computing systems are rather complex platforms supported by not only traditional sociology theory but also computer science and big data based applications. With the increase of the social computing systems' complexities, serious issues of social digital security and privacy have shown up since, in recent years, more and more social data leakage incidents are happening. This fact is due to reasons on many different aspects since there are many sources threatening the security and privacy of the social data in such a complex social computing system. In this paper, we improve the traditional social data protection schemes by combining the information fragmentation concepts with the distributed system architectures to build a novel social data protection scheme. We use social photo protection as the fundamental scenario and deploy our novel scheme to illustrate the improvement on the protection level with the protection analysis in detail. A security analysis of practically realizing such a scheme is also evaluated in this paper.

Keywords: Data security, social computing, big knowledge, selective encryption

Email address: xuan.chen@etu.sorbonne-universite.fr (Xuan CHEN)

1. Introduction

In the recent decade, with the development of information and communication technology, the social computing systems for social information processing through large social computing platforms have become more and more popular. Today, content social computing and sharing services have made social networking and platform providers immensely popular and important [3]. The social computing systems existed in our daily lives are more than just systems for only information exchanges. Now with the development of information technology, many previous social concepts and theories are able to develop on the social computing platforms. As shown in Figure 1, current social computing architecture is made up of mainly three layers and the most commonly used layer is formed of the social applications such as the social entertainments, social data service, and social communication [26].

Therefore, the social computing platform providers are not only the media for information exchange but also probably the most effective social data generation and storage system. Companies such as Facebook [18] or Twitter [32] could collect a huge amount of social data generated and transmitted every day and also get profits from this huge amount of social data. In such a process, the social data that supposed to be used for a form of social profile or self-expression are given commercial values and the result following is that the end-user's social data security and privacy are vulnerable.

In the past ten years, many web-based social network providers have attracted a huge number of people to join such as Facebook, Linkedin, and Twitter. Not only their real-time messages are shared from smart personal digital devices [31] through these social networks, some private information such as human profiles is also shared. These big social network companies are becoming big data-oriented service providers [30] since most social-related functions are free and the profits are coming from the data and knowledge mining from the social big data [20]. With this background, the data leakage incidents of the social data platforms are happening more frequently and more seriously. On

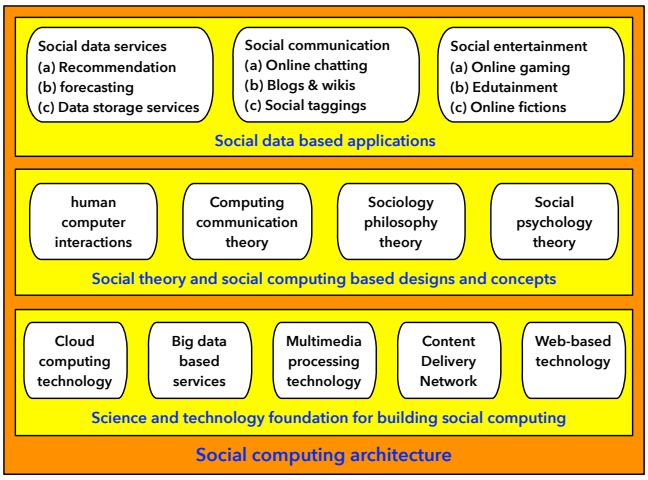

Figure 1: The complex modern social computing and communication systems [19] are made of different layers with theoretical and technical supports.

the one hand, social data service providers deal with a large number of external attacks. In 2018, a total of 1.5 million SingHealth patients' non-medical personal data were stolen from the health system in Singapore [11]. On the other hand, these social data service providers cannot be entirely trusted either. For instance, personal social data may be exploited in a malicious way such as in the Facebook and Cambridge Analytica data scandal which affected 87 million end-users in 2018 [21]. Moreover, there are other factors for privacy violation on social data such as the surveillance programs of the governments such as the PRISM program in 2013 [10].

Therefore, it is not totally fair to criticize the social platform providers for violating the end-users' privacy since social data privacy is not only violated by the one entities [1]. One of the classical and practical solutions is to protect social data before outsourcing it to the social data platform. Nowadays, there are many such new laws to force users to use such a standard. For instance, the European General Data Protection Regulation (GDPR) [24] states that the data owner should encrypt the data before outsourcing the data and also the key management should be protected by the data owner. Also, similarly, the

USA government recommends all digital data must be protected at rest, in transit, and in use, and the access to the data must be controlled [9]. In fact, these laws are transferring the data protection tasks to not only data service providers but also all the participates in the digital environment. The protection methodology is also transferred from the protection of the data to the protection of the key. However, in the social computing scenario, relying on the end-users to well manage their keys is also not practical because the reuse of the keys or passwords is frequently seen [23]. This will lead to a situation of key exposure situation as pointed in [29].

Therefore, since the existing methods seem not always sufficient for social data protection, we propose a new trend that can combine the protection with the data fragmentation [14]. The purpose of our design is to fragment the social data into different fragments and to use the key for the protection of only the important fragments. The transmission or processing of the social data will also be fragmented through different social computing platforms to achieve a higher level of protection against the situation such as the key exposure. Since the data are fragmented and the protection is performed selectively, the attackers must be able to acquire the key and the corresponding data fragments to decrypt which suppose to be transmitted or shared through different social computing providers. Since different social computing providers will come from different countries or be protected with different methods or laws, our method will increase the protection level for social data.

The main contribution of this paper includes: (1) the design of a secure social data outsourcing and sharing method in the current social big knowledge system; (2) extending the classical SE concept by designing a novel social data protection method by combining with the concept of fragmentation and distributed system while guaranteeing the efficiency of the proposed method.

We organize this paper following the order below. Section 2 gives several related works and the main research background of this paper. Section 3 illustrates the approach and the key algorithm of our proposed method. Section 4 gives a comprehensive security analysis and evaluation for the method with so-

cial data as examples. Section 5 discuss the method and future work. Finally, Section 6draws the conclusion of this study.

2. Research Background

In this section, we will list the research backgrounds including the current social platforms situation and the data privacy issues on these platforms. For the social platform illustration, we will use a brief case from Facebook [18] to illustrate the data-sharing cases in the current social networks. Then, the current data leakage incidents will be presented. The user behaviors are also one of the main reasons for the data leakage incidents and the background information will be listed in the third subsection as important background information.

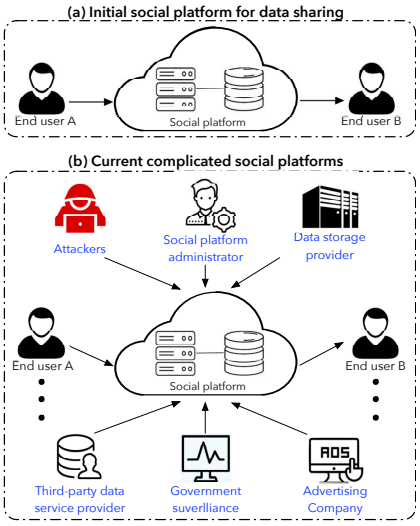

Figure 2: Current social platforms have been a complicated big knowledge system involved with many third-party entities: (a) initial designs of the social platforms are used for the brief communication for end-users; (b) nowadays, the social platforms have become a systems involving attackers, social platform admins, data storage providers, third-party data service providers, government surveillance, and advertising companies.

2.1. Social Platforms and Social Communications

The beginning of the social platform is based on the needs of the communication and information exchange between people. Based on the viewpoint of sociology, the social platform is the extension for the communication needs of multiple people [32]. However, in a long time of history, due to the limitation of the communication and information technology, there are only limited options for the communication and information exchange between people. In recent decades, with the invention of Internet technology, communication operations have changed and the possibility of a large group of social sensing and computing is realized.

For instance, one of such social computing system in the form of the microblogging [15] like Twitter, Jaiku, and Facebook. These kinds of systems allowed the users to broadcast brief text updates about small little things happening in their daily lives and work activities. This kind of social applications are welcomed today and have achieved popularity as an option for online social networking. For instance, according to [13], people like to update and share their daily life activities with friends, families, and co-workers which is privacy-related information. Also, they like to share their opinions on the information, news, and opinions on public affairs. Based on such facts, with the accumulation of such daily social data, platforms such as Twitter have become big data companies and the role of such platforms has become very important. Also, their functions have been becoming more and more complicated and there are many entities involved in the systems as shown in Figure 2. Nowadays, the social big data and the big knowledge based on such big data generation can be then built based on such complex social computing platforms such as Facebook.

2.2. Data Privacy Issues in Modern Social Computing Systems

As pointed in Section 1, on one hand, introducing the concept of big data and big knowledge for social computing brings advantages such as efficient, reliable, and economical services for the social computing users. However, since as pointed in Figure 2, there are many parties involved in such a system which

makes the data security especially privacy being challenged. As shown in many incidents such as the scandals in [4], directly bringing the existing security system into the complex social computing scenario cannot help solving all the challenges.

First, there are attackers trying to eavesdrop social data transmitted through the public communication channel which may compromise data transmitted in plain text. Then, due to the fact that the keys could be exposed in the complex social computing environment that many entities are involved, attackers may acquire the keys from one of the insecure or malicious entities inside the social computing environment. Moreover, since the data mining technology enabled the data service providers to benefit from understanding users' privacy information and sending advertisements, sometimes, the social computing providers become the threats [22]. The worse case is that the cloud service providers may directly hand over plain stored data for the authority's illegal surveillance without noticing the users [2].

With all the facts listed above, based on the end-user's viewpoint, the social computing service providers are not trustworthy so the protection taken at the end-user's end should be designed and implemented. Protecting social data before outsourcing to social computing providers is necessary to avoid directly attacking communication or any other operations on the social computing providers' ends to compromise end-users' data privacy. However, requiring end-users to just encrypt everything will introduce the issues of key management and cannot fundamentally solve the threats from attackers trying to acquire keys from key exposure situation. [17] Thus, the design must be able to avoid data leakage in the worst case assuming that an attacker has data stolen from social platform servers and has key leaked from other sources.

2.3. User Behaviors and Key Exposure

The research work shown in [29] is providing the anonymity for the end-user in the global mobility networks but the key exposure situation is not considered. The data could still suffer from the leakage of other communication entities

when there are malicious users in this environment expose the keys. However, the implementation is not performed considering the user behaviors such as repeated key usages.

However, since the computer and network security today is becoming a complex system with normally access revocation operations required and multiple parties involved, requiring all end-users to understand the security knowledge and manage well the sensitive information such as keys or passwords is not feasible. According to [23], the users may repeat use the same passwords for different systems to login which gives attackers the possibility to break multiple systems with one password. The main reason of reusing password is that most common end-users are lacking the knowledge of cryptography and are not aware of the weakness of using similar passwords will lead to threats such as dictionary attacks [25]. For the key usage, there will be similar situation that the keys set by end-users are sometimes based on the end-users' public information such as names or birthdays which could lead to the dictionary attack [22].

Therefore, we aim to consider such practical user behaviors, which are not helping for improving the data privacy schemes, and to build a method with privacy guaranteed even keys are exposed and some of the data are leaked as well. We believe by performing protection methods like combining encryption with fragmentation, the data protection will be enhanced although the keys and some parts of the social data might leak.

3. Architectures and Designs

In this section, the methods proposed in this paper will be presented. Firstly, we give a brief problem definition of the situation we want to solve in this paper. Then, we explain the system designs with the main algorithm and some necessary details. In the end, a brief case study will also be illustrated to explain how the proposed method works.

3.1. General Design

In this subsection, we illustrate the general design for the selective encryption procedures. As pointed in Section 1, there are two main threats we must deal with which are from the social platform-based communication process and the key exposure due to reasons like the user behaviors. Thus, the protection of the social data can be done by performing two steps including selective encryption and fragmentation. A brief example is shown in Figure 3, from the sender's viewpoint, the photo will be protected in the transformed domain with a selective encryption approach and then a part of selected data will be protected to be the "Protected small fragment" and the rest part will be re-transformed as the "Public large fragment". Then, the transmission can be done through two different social platforms and the receiver can recover the photos by reversing the protection process.

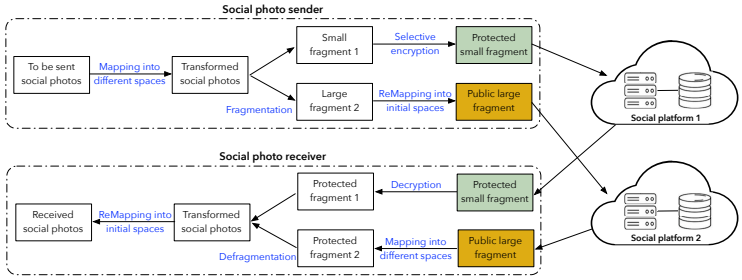

Figure 3: Conceptual architecture of the proposed social photo protection and transmission with selective encryption and fragmentation.

3.2. $Y'C_BC_R$ Space Representation for Images

$Y'C_BC_R$, is a set of color spaces used as a part of the color image pipeline in video and digital photography systems [5]. In fact, $Y'C_BC_R$ is made up of three components to represent the color images. Y' is the luminance component (light intensity is nonlinearly encoded based on gamma corrected RGB primaries) and C_B and C_R are the blue-difference and red-difference chroma components. According to [5], the math representation of the three components are listed in

equation 1:

$$\begin{aligned} Y' &= K_R \cdot R' + K_G \cdot G' + K_B \cdot B' \\ P_B &= \frac{1}{2} \cdot \frac{B' - Y'}{1 - K_B} \\ P_R &= \frac{1}{2} \cdot \frac{R' - Y'}{1 - K_R} \end{aligned} \tag{1}$$

where K_R, K_G, and K_B are ordinarily derived from the definition of the corresponding RGB space, and required to satisfy $K_R + K_G + K_B = 1$. The R', G' and B' are the normalized value of the color values of the three RGB layers, ranging from 0 to 1, and represents the level of the intensity in the corresponding color layers.

For the digital image case, when representing the signals in digital form, the results are scaled and rounded, and offsets are typically added. There are different values chosen for the K_R, K_G, and K_B. In this paper, the chose the ITU-R BT.601 (formerly CCIR 601) standard as the form to calculate the mapping between the RGB and the $Y'C_BC_R$ space. For the other standards, the results will be similar based on our brief tests which did not influence the final conclusions of this paper. The values for the K_R, K_G, and K_B are: $K_R = 0.299, K_G = 0.587, K_B = 0.114$. Thus, the digital $Y'C_BC_R$ (8 bits per sample), derived from the normalized R', G', and B' can be represented as in equation 2.

$$\begin{aligned} Y' &= 16 + (65.481 \cdot R' + 128.553 \cdot G' + 24.996 \cdot B') \\ C_B &= 128 + (-37.797 \cdot R' - 74.203 \cdot G' + 112 \cdot B') \\ C_R &= 128 + (112 \cdot R' - 93.786 \cdot G' - 18.214 \cdot B') \end{aligned} \tag{2}$$

Here, the prime symbols mean gamma correction is being used; thus R', G' and B' nominally range from 0 to 1, with 0 representing the minimum intensity (e.g., for the display of the color black) and 1 the maximum (e.g., for the display of the color white).

The resultant signals range from 16 to 235 for Y' (C_B and C_R range from 16 to 240). Therefore, this transform can be used for converting the RGB image into another space and the selective encryption scheme can be then operated.

The output of this transform is made up of three layers which include the Y' layer, C_B layer, and C_R layer. The Y' layer represents the image content in an approximate gray scale space while the C_B layer and C_R layer represent the details for the color elements with an example shown in Figure 4. The color Lenna image can be represented as three layers of the red layer, green layer, and blue layer as shown in Figure 4-(a), (b), and (c), respectively. Then these three layers can be converted into the $Y'C_BC_R$ space representation as the Y', C_B, and C_R layers as shown in Figure 4-(d), (e), and (f), respectively.

3.3. Algorithm presentation for the proposed scheme

The system architecture is shown in Figure 3. In this subsection, we will use the algorithm presentation to illustrate how exactly the protection is performed based on the $Y'C_BC_R$ space representation. Firstly, as mentioned in Section 2, the protection on the different layers should be different as well according to the relative importance levels. According to the physical definition of the $Y'C_BC_R$ space representation, the Y' layer represents the basic image contents while the other two layers carry the details for rebuilding the color elements. Thus, in this paper, we propose to protect more ratios of the Y' layer while protecting fewer ratios of the C_B and C_R layers. The notations and definitions used in this paper are shown in Table 1. The protection ratio α is an integer with the range from 1 to 8 and particularly means how many bits of one byte will be protected in this method. For the discussion on the protection ratio, we evaluate different values in Section 4.

In this paper, we explore the selective encryption on the different layers of the $Y'C_BC_R$ space representation for social images with different protection levels. The protection can be selectively performed in the $Y'C_BC_R$ space and then the protected results will be converted back to the RGB space and will form the final protected photos in the end.

The algorithm presentation for the selective protection details is shown in Algorithm 1. Firstly, the input social photo P will be represented into the RGB space which is $P = \{P_R, P_G, P_B\}$. Then, a converation from the RGB

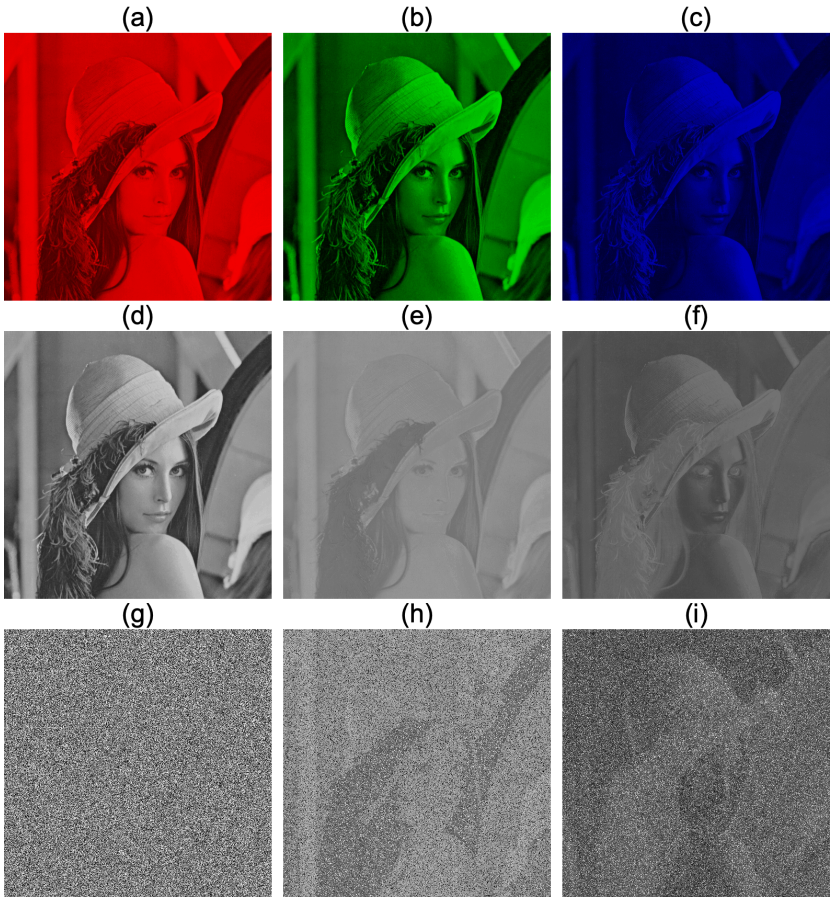

Figure 4: An example for the converting from the color image to the $Y'C_BC_R$ space representation: (a), (b), (c) the Red, Green, and Blue representation for the Lenna image; (d) (e) (f) the Y', C_B, C_R representation converted from the RGB representation of the Lenna image; (g) protected Y' layer; (h); protected C_B layer; (i)protected C_R layer.

Table 1: Major notations used in the algorithm and their definitions.

Notation	Definition
P	Input social photo
P_R	Red layer representation
P_G	Green layer representation
P_B	Blue layer representation
$P_{Y'}$	Y' layer representation
P_{CB}	C_B layer representation
P_{CR}	C_R layer representation
$width$	Width of the input photo
$length$	Length of the input photo
K	Key used for protection
α	Protection level factor
$PRNG$	Pseudorandom number generator
$P_{private}$	Output public fragment
P_{public}	Output private fragment

space to the $Y'C_BC_R$ space will be performed to get the representation as $\{P_{Y'}, P_{CB}, P_{CR}\}$. This space is the one we will operate the selection and lightweight protection. In fact, we initially set the protection level factor $\alpha = 4$ as the default protection level for the important layer which is the $P_{Y'}$ layer. The definition of α is how many bits out of one byte will be selectively protected for the important $P_{Y'}$ layer. Accordingly, the protection on the other two layers (P_{CB}, P_{CR}) will be relatively lightweight since we will pick one bit out of one byte to protect. The factors such as α can be tuned to adjust the protection levels of the photos and in this paper, we will illustrate the protection methods with setting $\alpha = 4$ and protect only 1 bit out of one byte in the P_{CB} and P_{CR} layers.

Since our initial motivation is to design selective encryption in a lightweight manner and the protection must be able to combine with the separation between the protected parts and the unprotected parts. Besides the separation factor α, we also use a key K as the seed for the generation of the selection process. For each photo, the K will be used to generate a sequence of the indexes with a Pseudo Random Number Generator (PRNG). According to [16], the PRNG

is a generator that can generate the same sequence of numbers under the condition that only the same seed is used. Thus, based on an attacker's viewpoint, the selective encryption can be performed on a sequence of randomly selected bits out of every byte with the protected bits are mixed with the unprotected bits. In fact, based on the viewpoint of fragmentation, the selected bits can be accumulated and generate a fragment that is more important than the other ones. For further protection, this fragment can be encrypted with standard encryption schemes such as AES [6]. The other fragment that is considered as less important compared with this one and can be protected in a lightweight manner or left in plaintext.

The details are listed in Algorithm 1 and the output of the algorithm is two data fragment $P_{private}$ and P_{public}. The total process of this algorithm can be then represented as: $P = P_{private} + P_{public}$. Considering the protection level, the $P_{private}$ can be the same protection level compared with the standard encryption schemes while the P_{public} will be left in plain in this paper. Later in Section 4, we will illustrate the protection levels of different fragments to evaluate the protection level in different metrics. The other factor should be considered in about the storage space usage by deploying this selective protection scheme. If without the compression process, the storage space taken by P_{public} will be the same with P and the $P_{private}$ will be 20.8% of the P. Totally, the transmission of P will be increased with 20.8%. However, if there is compression step for the RGB photo such as JPEG compression, the ratio will be changed according to the compression standard used The evaluation of the storage space usage will be given in Section 4.

4. Results and Evaluations

In this section, different security tests on the proposed method are performed in order to evaluate its protection levels. The private fragment $P_{private}$ of one social photo is supposed to be protected with standard encryption schemes such as AES-128 [6]. Of course, this encryption scheme can be easily replaced with

Algorithm 1: Algorithm of the selective encryption.

Input: Social photo P and Key K.

Output: Public fragment P_{public} and private fragment $P_{private}$.

$\{P_R, P_G, P_B\} = P$; /*Read input photo P and convert it as three RGB layers.*/

$\{P_{Y'}, P_{CB}, P_{CR}\} = \{P_R, P_G, P_B\}$; /*Convert the RGB representation of the photo p into the $Y'C_BC_R$ space.*/

for $w \leftarrow 1$ to $width$ **do**

for $l \leftarrow 1$ to $length$ **do** /*Protect pixel by pixel with two loops.*/

for $i \leftarrow 1$ to α **do**

$Index = PRNG(K + i, w, l)$; /*Generate index for the selective protection on $P_{Y'}$.*/

$tempBinPY' = dec2bin(P_{Y'}(w, l))$; /*Transform the $P_{Y'}$ into binary representation.*/

$tempBinPY'(Index) =$ **NOT** $tempBinPY'(Index)$; /*Protect the selected bits in $P_{Y'}$ with NOT operation.*/

$P'_{Y'} = bin2dec(tempBinPY')$; /*Transform the protected bit strings back to integer representation.*/

$P_{private} = P_{private} + tempBinPY'(Index)$; /*Accumulate the private fragment.*/

end for

$IndexBR = PRNG(K, w, l)$; /*Generate index for the selective protection on P_{CB}, P_{CR}.*/

$tempBinPCB = dec2bin(P_{CB}(w, l))$; /*Transform the $P_{Y'}$ into binary representation.*/

$tempBinPCR = dec2bin(P_{CR}(w, l))$; /*Transform the $P_{Y'}$ into binary representation.*/

$tempBinPCR(Index) =$ **NOT** $tempBinPCR(Index)$; /*Protect the selected bits in P_{CR} with NOT operation.*/

$tempBinPCB(Index) =$ **NOT** $tempBinPCB(Index)$; /*Protect the selected bits in P_{CB} with NOT operation.*/

$P'_{CR}(w, l) = bin2dec(tempBinPCR)$;

$P'_{CB}(w, l) = bin2dec(tempBinPCB)$; /*Transform the protected bit strings back to integer values.*/

$P_{private} = P_{private} + tempBinPCB(Index) + tempBinPCR(Index)$; /*Accumulate the private fragment.*/

end for

end for

$\{P'_R, P'_G, P'_B\} = \{P'_{Y'}, P'_{CB}, P'_{CR}\}$; /*Convert the protected representation back to the RGB space.*/

$P_{public} = \{P'_R, P'_G, P'_B\}$; /*Rebuild the photo with protected results*/

Return $P_{private}$, P_{public};

any other encryption algorithms as the flexibility always remains. Thus, the security property of the private fragment $P_{private}$ will not be analyzed in this paper. Therefore, the challenge turns to be the possibility of recovering the social photo P only from the P_{public}. More specifically, the question can then be summarized as that an attacker could get the P_{public} which is in plaintext transmitted through the social computing platforms. This P_{public} will be used to recover the original social photo P at the attacker's end. Therefore, in this section, we will only show the analysis results of the P_{public}.

4.1. Visual Effects Evaluation

In this subsection, we only evaluate the protection results on visual effects since the visual results are the most effective way to evaluate the protection levels. As pointed in Section 3, the selection ratio α for the encryption is not used for the most important layer Y' which is representing the basic image elements as shown in Figure 5 (c). Since our initial purpose is to propose lightweight fashion protection, the other two layers are processed with only one bit selectively protected. Then, we compare the original social photo and the protected results as shown in Figure 5 (a) and (b), respectively. Based on the observation, the protection for the Y' layer is relatively better than the final protection on the input social photo P.

Figure 5: Visual results comparison between: (a) original social photo P; (b) protected social photo P.

Also, we particularly test the different selection ratio of α to understand the selection situation for the protection. In fact, the minimum protection ratio is

when $\alpha = 1$ which means only 1 bit out of each pixel in the Y' layer is encrypted such that the protection level is minimum as well. Since in the initial design, the protection ratio only works on the important part of the photo which is the Y' layer of the three layers. In order to test the protection selection ratio α, we show the visual effects of the protection for all the possible values for the selection ratio of α with an example shown in Figure 6. In fact, since the practical protection method in this paper is to do the XOR operation for the selected bits, the protection level will achieve the best when half of the bits are selected to perform the XOR operation. If the α is more than 4 which means more than half of the bits will be XORed which is actually the same with the condition of smaller α value. If $\alpha = 8$, the result we get in Figure 6 (i) is the same photo with the input but with only all the pixels are minus by 255. Therefore, for the important Y' layer, the selection ratio α will be set as 4 to achieve the maximum protection and for the other two layers, the selection ratio α will be set as 1 to reduce the execution time and the footprint of the $P_{private}$.

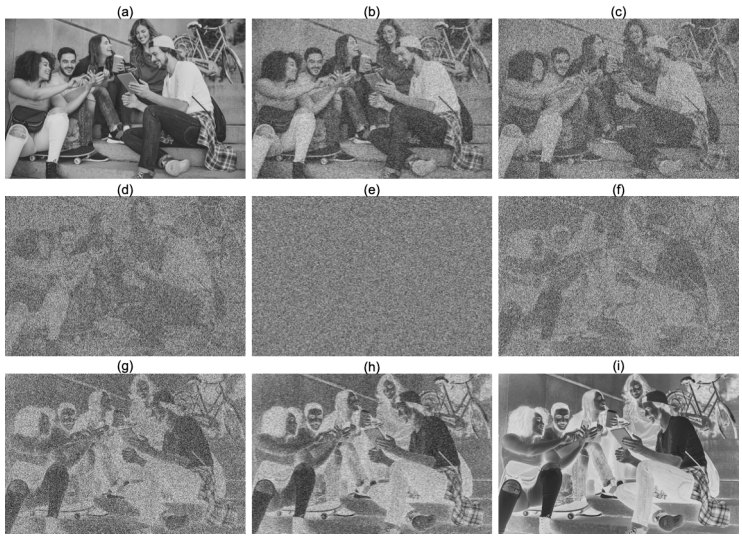

Figure 6: Comparison of visual results of protection on the Y' layer with different selection ratio α: (a) initial social photo P; (b) $\alpha = 1$; (c) $\alpha = 2$; (d) $\alpha = 3$; (e) $\alpha = 4$; (f) $\alpha = 5$; (g) $\alpha = 6$; (h) $\alpha = 7$; (i) $\alpha = 8$.

Therefore, based on the observation of the visual effects for images, the pro-

posed method can achieve a hard visual degradation which can protect the image contents with a lightweight manner. Also, in order to measure the statistical results on the protection, Peak Signal-to-Noise Ratio (PSNR) [12] is used to quantify the visual degradation compared with the original input photo. PSNR is derived from the Mean Squared Error (MSE), while MSE represents the cumulative squared error between two matrices. A low PSNR value [12] indicates that there is a high difference between the original photo and the protected photo. In this section, we tested different social photos to show the protection level considering the PSNR values and the results are shown in Figure 7 (a). Since in this test, the selection ratio α is set to 4 but the tests on three color layers are shown in 7 (b).

As shown in Figure 7, the PSNR value for different social photos and the results show that the PSNR values are always less than 12dB. For the three color layers, the PSNR values are similar to the values of the PSNR of the photos which indicate that all the three color layers are very different compared with the initial photos.

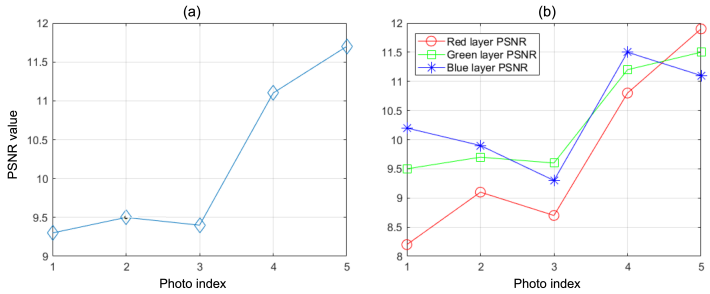

Figure 7: PSNR test for five social photos: (a) comparison of color photos' PSNR between the initial social photos and the protected social photos; (b) comparison of PSNR values for three layers (RGB layers) respectively.

4.2. Uniformity and Correlation Analysis

According to [27], the analysis of evaluating the protection should also be performed to prove this scheme can protect against the statistical attacks. In order to validate the protection of this method, different statistical security tests

are being applied with independence between the input photos P and the P_{public} of the input photo P. According to the related work shown in [27], we perform the comparison of the protection as follows: red layer value histogram, green layer value histogram, and blue layer value histogram of input social photo P as shown in Figure 8 (a), (b), and (c) respectively, with the red layer value histogram, green layer value histogram, and blue layer value histogram of the P_{public} of the input photo P as shown in Figure 8 (a), (b), and (c), respectively.

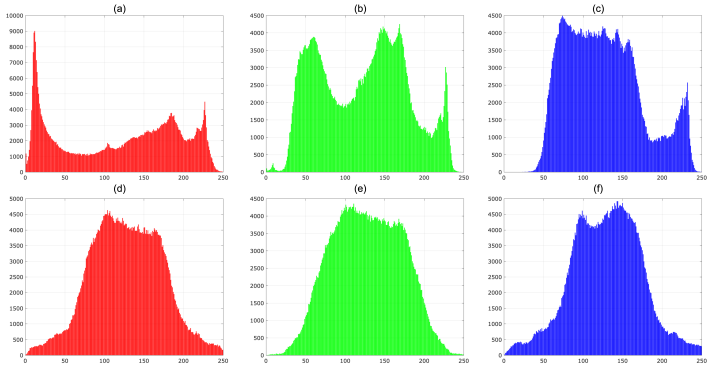

Figure 8: Histogram of the RGB layer comparison between the initial photo and the P_{public} of this photo: (a) red layer value histogram of input social photo; (b) green layer value histogram of input social photo; (c) blue layer value histogram of input social photo; (d) red layer value histogram of P_{public} of input social photo; (e) green layer value histogram of P_{public} of input social photo; (f) blue layer value histogram of P_{public} of input social photo.

In order to measure the quantitative results of the protection, we will use the 2D-correlation to measure the similarity of the initial input social photo P and the protected and public results P_{public}. Normally, the 2D-correlation is a measure of similarity of two series as a function of the displacement of one relative to the other.

Two-dimensional correlation analysis is a mathematical technique that is used to study changes in measured signals. 2D correlation analysis is used for its main advantage: increasing spectral resolution by spreading overlapping peaks over two dimensions and as a result simplification of the interpretation of one-dimensional spectra that are otherwise visually indistinguishable from each

other. Here, in this paper, we use the 2D-correlation to measure the features between two potential similar signals. If we set the two dimensional signals of photo P and P_{public} as A and B, the correlation factor for 2D is defined as:

$$r = \frac{\sum_m \sum_n (A_{mn} - \bar{A})(B_{mn} - \bar{B})}{\sqrt{\sum_m \sum_n (A_{mn} - \bar{A})^2 \sum_m \sum_n (B_{mn} - \bar{B})^2}} \tag{3}$$

where $\bar{A} = mean_{2d}(A)$ and $\bar{B} = mean_{2d}(B)$. The correlation coefficient, denoted by r, is a measure of the strength of the straight-line or linear relationship between two variables. The well-known correlation coefficient is often misused because its linearity assumption is not tested. The output value r is a value in the interval between +1 and -1, including the end values +1 or -1. If $r = 0$, that means there is no linear relationship between the two variables. The fact that $r = +1$ indicates a perfect positive linear relationship as one variable increases in its values, the other variable also increases in its values through an exact linear rule. The fact that $r = -1$ indicates a perfect negative linear relationship – as one variable increases in its values, the other variable decreases in its values through an exact linear rule.

In this paper, we evaluate the r^2 to indicate the similarity between the initial photo P and the protected public photo P_{public}. The comparison of the 2D-correlation r^2 in this paper will be done for three color layers for the different social photos. The experiments for different social photos are shown in Figure 9 for three color layers and the Squared correlation factors (r^2) proves the low correlation between the initial social photo P and $P_{private}$ on three color layers respectively.

4.3. Cryptoanalysis and Discussions

In this subsection, typical published cryptanalytic cases are considered and brief analysis of the proposed method against several cryptanalytic attacks is provided from a cryptanalysis viewpoint. The proposed method is considered to be public and the attacker has complete knowledge of all steps but no knowledge about the secret key. In this method, we propose that even with the key

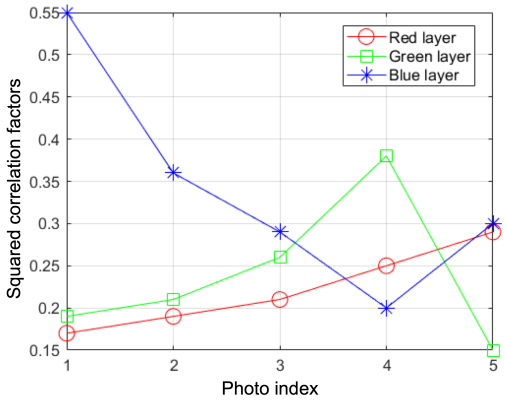

Figure 9: Correlation factors for the different social photos.

exposure, the method can still protect the social data as long as not all the data fragments are acquired by the attackers with the key expose.

For the private fragment $P_{private}$ of the social photo P, we use the AES to encrypt the data fragments since the AES has options for keyspace that can be 2^{256} [6], which is sufficiently large to make the brute-force attack almost infeasible. Furthermore, differential and linear attacks for the $P_{private}$ would also become ineffective. For the public fragment P_{public}, the data that can be used for the rebuilding is very limited according to the visual and statistical analysis. The spatial redundancy between the public fragment P_{public} of initial social photo P is removed and a high randomness degree of all fragments is achieved. Different statistical tests such as the histogram analysis and correlation analysis are applied to validate the independence and uniformity property. Consequently, these results indicate that no useful information can be detected from the public fragment P_{public} to rebuild the initial social photo P. This validates the robustness of the proposed method and their high resistance to statistical attacks.

For further cases like a single plain-text failure, Initialization Vector (IV) [7] or counters could be introduced to generate dynamic keys for each of the chunk. In such a case, we increase the difficulty levels for an attacker to recover the

dynamic keys that are changed for every input chunk. For accidental key disclosure, the attacker must be able to retrieve all the fragments to recover the input social photo P due to the relations we build among the different fragments. For the worst case, we assume an end-user is using the same key all the time and this key has been leaked to the attacker, and the attacker managed to get all fragments of social photo P_k to recover the initial social photo P_k. However, for the other social photos such as P_{k+1}, the attacker will still need to steal all the fragments for the P_{k+1} and the knowledge to the key will not help to recover other social photos.

4.4. Discussion on Comparison

In this paper, we adopted the concept of data fragmentation with the full encryption to design a novel lightweight encryption algorithm for the social photos in the modern social big knowledge systems. The purpose is not to replace the existing full encryption algorithms such as AES or RSA. The fact is that normally all the lightweight encryption approaches are less secure than the standard encryption algorithms such as AES but the lightweight encryption algorithms could win on the performance aspects. There are some other lightweight encryption approaches such as Selective Encryption (SE) could prove the efficiency advantage by comparing the computing tasks to be performed. The approach proposed in this paper will perform only the protection process based on the AES but there is only less than 20.8% and there are almost no other additional calculation tasks expect the XOR operations.

5. Discussion and Future Work

In this paper, we proposed a selective encryption method to selectively protect social data sharing with a practical example of social photos in the current social big knowledge system. We proposed to combine the fragmentation with the selective encryption in a lightweight manner. By applying this approach, this method can be used to resist the situation that the key is leaked and the

social computing service provider is not trustworthy. Fragmentation of storage could provide the necessary protection over the different fragments which is the additional protection for the social data. Compared with the existing full encryption algorithms such as AES or RSA which just transfers the protection from the data to the protection on the keys, the fragmentation based protection could provide an additional protection against the situations such as key leakage. Assuming a very simple use case that a user protects the social data stored on the clouds with AES and a simple key, the cloud providers may be able to decrypt the stored data by guessing the key with dictionary attacks. However, if the social data is protected with fragmentation methods, the social computing providers will not be able to acquire other data fragments from the other social computing providers which further prevents the privacy violation.

There are still two other issues remaining for the future work. One important issue for selective encryption methods is still needed to be investigated which is how to improve the importance levels of the different data fragments. According to the recent research works [28], the deep learning techniques are already deployed on the smartphone scenarios which may also be used to determine the better selection methods. Thus, one future direction is to use the deep learning method to improve the selection steps in the protection method. Also, we will deploy this scheme with a more practical environment such as a multiple-party communication scenario [8]. The practical research direction should be to research the possibility to define different cloud servers with different trust levels and disperse all fragments. The performance should be investigated further with practical social computing environments such as the PC or smartphones to get real-world results.

6. Conclusion

In this paper, we presented an efficient social data protection method for the social big knowledge system. The purpose is to solve the cyber threats to data security and privacy in the social computing scenario considering end-

users' behaviors that keys might be reused and leaked from other sources. A SE algorithm combined with fragmentation and dispersion for storage is designed to protect data even when both the key and the public fragment of social photos are leaked. We used different test methods to prove the effectiveness of our proposal and the protection can be achieved with only 20.8% of the initial data protected. Therefore, we proposed an approach for securely sharing and transmitting the social data that can avoid data recovery even with both key and public fragments compromised.

References

[1] Md Liakat Ali, John V Monaco, Charles C Tappert, and Meikang Qiu. Keystroke biometric systems for user authentication. *Journal of Signal Processing Systems*, 86(2-3):175–190, 2017.

[2] James Ball. Nsa's prism surveillance program: how it works and what it can do. *The Guardian*, 8, 2013.

[3] Gema Bello-Orgaz, Jason J Jung, and David Camacho. Social big data: Recent achievements and new challenges. *Information Fusion*, 28:45–59, 2016.

[4] Hal Berghel. Equifax and the latest round of identity theft roulette. *Computer*, 50(12):72–76, 2017.

[5] Nadia Brancati, Giuseppe De Pietro, Maria Frucci, and Luigi Gallo. Human skin detection through correlation rules between the ycb and ycr subspaces based on dynamic color clustering. *Computer Vision and Image Understanding*, 155:33–42, 2017.

[6] Joan Daemen and Vincent Rijmen. *The design of Rijndael: AES-the advanced encryption standard*. Springer Science & Business Media, 2013.

[7] Maththaiya Durai and Behnam S Arad. A pipelined implementation of hash stream1-synthetic initialization vector encryption algorithm. *International Journal for Computers & Their Applications*, 23(2), 2016.

[8] Chong-zhi Gao, Qiong Cheng, Xuan Li, and Shi-bing Xia. Cloud-assisted privacy-preserving profile-matching scheme under multiple keys in mobile social network. *Cluster Computing*, 22(1):1655–1663, 2019.

[9] Timothy Grance and Wayne Jansen. Guidelines on security and privacy in public cloud computing. Technical report, 2011.

[10] Glenn Greenwald and Ewen MacAskill. Nsa prism program taps into user data of apple, google and others. *The Guardian*, 7(6):1–43, 2013.

[11] Steve Hambleton et al. A glimpse of 21st century care. *Australian Journal of General Practice*, 47(10):670–673, 2018.

[12] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of PSNR in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008.

[13] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. AcM, 2010.

[14] Marc Lacoste, Markus Miettinen, Nuno Neves, Fernando MV Ramos, Marko Vukolic, Fabien Charmet, Reda Yaich, Krzysztof Oborzynski, Gitesh Vernekar, and Paulo Sousa. User-centric security and dependability in the clouds-of-clouds. *IEEE Cloud Computing*, 3(5):64–75, 2016.

[15] Lara Lomicka. Twitter and micro-blogging and language education. *Language, Education and Technology*, pages 1–12, 2017.

[16] MA Murillo-Escobar, C Cruz-Hernández, L Cardoza-Avendaño, and R Méndez-Ramírez. A novel pseudorandom number generator based on pseudorandomly enhanced logistic map. *Nonlinear Dynamics*, 87(1):407–425, 2017.

[17] SA O'Brien. Giant equifax data breach: 143 million people could be affected. *CNN Tech*, 2017.

[18] Tiffany A Pempek, Yevdokiya A Yermolayeva, and Sandra L Calvert. College students' social networking experiences on facebook. *Journal of applied developmental psychology*, 30(3):227–238, 2009.

[19] Kay Peters, Yubo Chen, Andreas M Kaplan, Björn Ognibeni, and Koen Pauwels. Social media metrics—a framework and guidelines for managing social media. *Journal of interactive marketing*, 27(4):281–298, 2013.

[20] Matthew A Russell. *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More.* " O'Reilly Media, Inc.", 2013.

[21] Olivia Solon and Oliver Laughland. Cambridge analytica closing after facebook data harvesting scandal. *The Guardian*, 2018.

[22] Jeffrey M Stanton, Kathryn R Stam, Paul Mastrangelo, and Jeffrey Jolton. Analysis of end user security behaviors. *Computers & security*, 24(2):124–133, 2005.

[23] Hung-Min Sun, Yao-Hsin Chen, and Yue-Hsun Lin. opass: A user authentication protocol resistant to password stealing and password reuse attacks. *IEEE Transactions on Information Forensics and Security*, 7(2):651–663, 2011.

[24] Paul Voigt and Axel Von dem Bussche. The EU general data protection regulation (GDPR). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.

[25] Ding Wang and Ping Wang. Offline dictionary attack on password authentication schemes using smart cards. In *Information security*, pages 221–237. Springer, 2015.

[26] Peng Wang, BaoWen Xu, YuRong Wu, and XiaoYu Zhou. Link prediction in social networks: the state-of-the-art. *Science China Information Sciences*, 58(1):1–38, 2015.

[27] Shujiang Xu, Yinglong Wang, Jizhi Wang, and Min Tian. Cryptanalysis of two chaotic image encryption schemes based on permutation and XOR operations. In *Computational Intelligence and Security, 2008. CIS'08. International Conference on*, volume 2, pages 433–437. IEEE, 2008.

[28] Qingshui Xue, Haozhi Zhu, Xingzhong Ju, Haojin Zhu, Fengying Li, Xiangwei Zheng, and Baochuan Zuo. A video-selection-encryption privacy protection scheme based on machine learning in smart home environment. In *International Conference on Artificial Intelligence for Communications and Networks*, pages 65–76. Springer, 2019.

[29] Jia Yu, Kui Ren, Cong Wang, and Vijay Varadharajan. Enabling cloud storage auditing with key-exposure resistance. *IEEE Transactions on Information forensics and security*, 10(6):1167–1179, 2015.

[30] Cheng Zhang, Zhi Liu, Bo Gu, Kyoko Yamori, and Yoshiaki Tanaka. A deep reinforcement learning based approach for cost-and energy-aware multi-flow mobile data offloading. *IEICE Transactions on Communications*, page 2017CQP0014, 2018.

[31] Cheng Zhang and Zixuan Zheng. Task migration for mobile edge computing using deep reinforcement learning. *Future Generation Computer Systems*, 96:111–118, 2019.

[32] Dejin Zhao and Mary Beth Rosson. How and why people Twitter: the role that micro-blogging plays in informal communication at work. In *Proceedings of the ACM 2009 international conference on Supporting group work*, pages 243–252. ACM, 2009.