Compressed Sensing Based Encryption Approach for Tax Forms Data

Adrian Brezulianu
"Gheorghe Asachi" Technical University of Iasi
Iasi, Romania

Monica Fira
Romanian Academy
Institute of Computer Science
Iasi, Romania

Marius Daniel Peştină
"Gheorghe Asachi" Technical University of Iasi
Iasi, Romania

***Abstract*—In this work we investigate the possibility to use the measurement matrices from compressed sensing as secret key to encrypt / decrypt signals. Practical results and a comparison between BP (basis pursuit) and OMP (orthogonal matching pursuit) decryption algorithms are presented. To test our method, we used 10 text messages (10 different tax forms) and we generated 10 random matrices and for distortion validate we used the PRD (the percentage root-mean-square difference), its normalized version (PRDN) measures and NMSE (normalized mean square error). From the practical results we found that the time for BP algorithm is much higher than for OMP algorithm and the errors are smaller and should be noted that the OMP does not guarantee the convergence of the algorithm. We found that it is more advantageous, for tax forms (or other templates that show no interest for encryption) to encrypt only the recorded data. The time required for decoding is significantly lower than the decryption for the entire form**

Keywords—compressed sensing; encryption; security; greedy algorithms

I. INTRODUCTION

The theory of compressed sensing, perfected in the past few years by prestigious researchers such as D. Donoho [1], E. Candès [2], M. Elad [3], demonstrates the feasibility of recovering sparse signals from a number of linear measurements, dependent with the signal sparsity. Compressed sensing (CS) is a new method which draws the attention of many researchers and it is considered to have an enormous potential, with multiple implications and applications, in all fields of exact sciences [1-4]. Specifically, CS is a new technique for finding sparse solutions to underdetermined linear systems. In the signal processing domain, the compressed sensing technic is the process of acquiring and reconstructing a signal that is supposed to be sparse or compressible.

The perfect secrecy together with the secret communication is a well-defined field of research, being a difficult problem in the domain of information theory. One of the requirements for the information theoretic secrecy is to assure that a spy who listens a transmission containing messages will collect only small number of information bits from message. Additionally, it should provide protection against of an computationally unlimited adversary based on the statistical properties of a system. Shannon introduced the idea of perfect secrecy, in his fundamental paper [5].

An encryption idea by utilizing CS has been mentioned for the first time in [7], but not been addressed in detail [6]. In paper [8], the secrecy of CS is researched, and whose result is that CS can provide a computational guarantee of secrecy. In [9] examine the security and robustness of the CS-based encryption method. In paper [10], the authors describe a new coding scheme for secure image using the principles of compressed sensing (CS) and they analyze the secrecy of the scheme.

II. BACKGROUND

A. Compressed Sensing

Compressed sensing studies the possibility of reconstructing a signal x from a few linear projections, also called measurements, given the a priori information that the signal is sparse or compressible in some known basis Ψ.

To define sparsity precisely, we introduce the following notation: for Ψ - a matrix whose columns form an orthonormal basis, we define a K-sparse vector $x \in R^n$ as $x = \Psi\theta$, where $\theta \in R^N$ has K non-zero entries (i.e., is K-sparse) and Ω_K as the set of K indices over which the vector θ is non-zero.

The vectors on which x is projected onto are arranged as the rows of a nxN projection matrix Φ, n < N, where N is the size of x and n is the number of measurements. Denoting the measurement vector as y, the acquisition process can be described as:

$$y = \Phi x = \Phi\Psi\gamma \quad (1)$$

$$\hat{\gamma} = \arg\min_{\gamma} \|\gamma\|_{l_0} \quad subject\ to \quad y = \Phi\Psi\gamma \quad (2)$$

$$\hat{x} = \Psi\hat{\gamma} \quad (3)$$

The equations system (1) is obviously undetermined. Under certain assumptions on Φ and Ψ, however, the original expansion vector γ can be reconstructed as the unique solution to the optimization problem (2); the signal is then reconstructed with (3). Note that (2) amounts to finding the sparsest decomposition of the measurement vector y in the dictionary $\Phi\Psi$. Unfortunately, (2) is combinatorial and unstable when considering noise or approximately sparse signals.

For a K-sparse signal, only "K+1 projections of the signal onto the incoherent basis are required to reconstruct the signal with high probability"[5]. In this case, is necessary to use combinatorial search with huge complexity. In [1] and [2] is proposed tractable recovery procedures based on linear programming. In these papers is demonstrated that the tractable recovery procedures obtain the same results toward combinatorial search when for signal reconstruction are used aprox. 3 or 4 cK projections.

Two directions have emerged to circumvent these problems:

- Pursuit and thresholding algorithms seek a sub-optimal solution of (2)
- The Basis Pursuit algorithm [1] relaxes the l_0 minimization to, solving the convex optimization problem (4) instead of the original.

$$\hat{\gamma} = \arg\min_{\gamma} \|\gamma\|_{l_1} \quad subject\ to \quad y = \Phi\Psi\gamma \qquad (4)$$

The matrix Φ satisfies a restricted isometry property of order K whether there is a constant $\delta_K \in (0,1)$ such that the inequation (5),

$$(1-\delta_K)\|x\|_2^2 \le \|\Phi x\|_2^2 \le (1+\delta_K)\|x\|_2^2 \qquad (5)$$

holds for all x with sparsity K.

A. Notions of secrecy and Model

In cryptography, "a secret key system is an encryption system where both sender and receiver use the same key to encrypt and respectively, decrypt the message" [11-12].

A conventional encryption scheme consists of five elements [13-14]:

• *Plain text:* This is the original message or input information for the encryption algorithm.

• *Encryption algorithm*: This algorithm performs various substitutions and modifications to the clear text.

• *Secret Key:* This key is an input to the encryption algorithm.

• *Ciphertext:* The text resulting from encryption algorithm and it is depends on the plaintext and the secret key. Thus, for a given message, two different secret keys produce two different ciphertexts.

• *Decryption algorithm:* This algorithm is the inverse of the encryption algorithm. The decryption algorithm is applied with the same secret key to the ciphertext in order to get the original clear text.

Following two elements must be taken into account in order to achieve a secure encryption [15]:

1) The encryption algorithm should be very strong. If an attacker knows the encryption algorithm (encryption) and has access to one or more ciphertext, he cannot decrypt the ciphertext or find the secret key.

2) Both the transmitter and the receiver must obtain the secret key in a safe manner (on a secure communication channel) and to keep it secret.

Based on previous remarks, in Figure 1 (in the upper half) is shown the basic model for CS and it includes two major aspects: measurements taking and signal recovery. The measurements taking involve an encryption algorithm and signal recovery is associated with a decryption algorithm from the perspective of symmetric-key cipher. The relationship between CS and symmetric cryptography indicates that some possible cryptographic features can be embedded in CS.

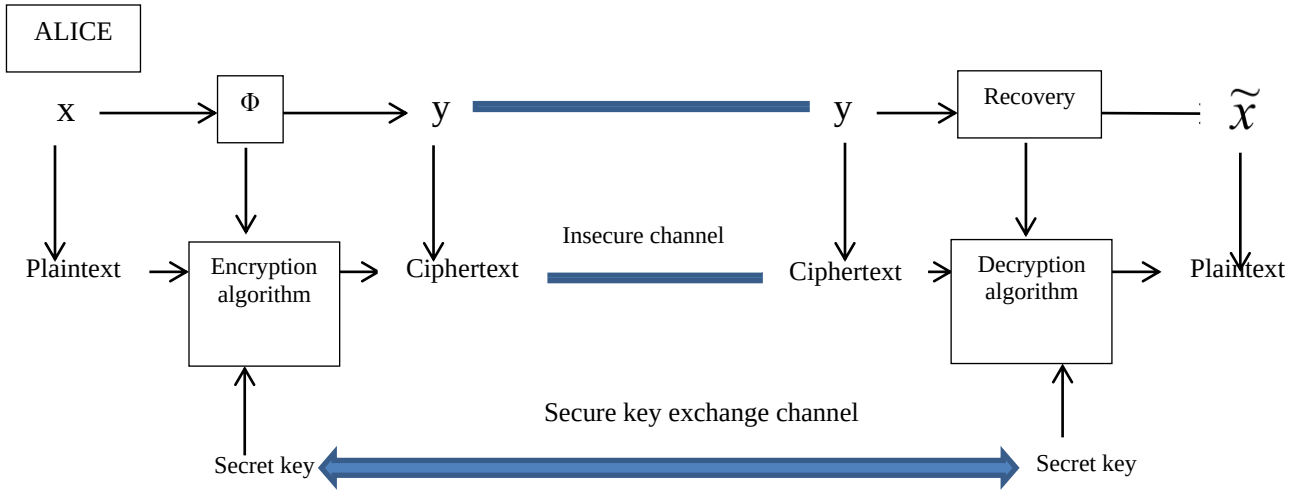

Fig. 1. The relationship between CS and symmetric-key cipher

The classical example of communication of a secret message from Alice to Bob assumes that Alice must use key from the set of keys. In this paper, let be i a key chosen by Alice with equal probability, and used to encrypt the message x with help of Φ_i matrix (via matrix multiplication operation). The result of multiplication is the cryptogram y which is

Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors*).

transmitted to Bob. The recipient knows the key used for encryption of the message. Knowing Φ_i and y, the compressed sensing literature provides conditions for x and Φ_i to allow the recovery of the original message x. The classical example of secret message communication assumes that the Alice's encrypted message y is being intercepted by an eavesdropper named Eve. For the third person, the used key the message encryption is unknown.

In our case, the measurement matrix Φ can be selected from a set of keys that is known for the transmitter (Alice) and the permitted receiver (Bob). Each random measurement matrix Φ is generated with a seed which can be exchanged through a secure approach between two desired sides [16-17].

A computational encryption scheme is secure if the ciphertext has one or two properties:

- The cost of breaking ciphertext is much higher than the encrypted information.
- The time needed for breaking ciphertext is longer than the lifetime of the information.

A brute force attack on the compressive sampling based encryption scheme would be guessing the linear measurement matrix Φ_i. Thus, an eavesdropper, e.g. Eve, could directly try to do this by performing an exhaustive search over a "grid" of values for Φ_i. But, the step size of this grid is critical because a too large step size may cause the search to miss the correct value and a too small grid size will increase the computational task unnecessarily.

The computational cost of signal reconstruction is high. For the best optimization algorithm (BP), the computational cost is in the order of $O(N^3)$ and a random search will make the search too expensive.

III. Simulations and Discussions

To test our method, we used 10 text messages (10 different tax forms) and we generated 10 random matrices.

To validate the decoding results, we evaluate the distortion between the original plaintext and the reconstructed plaintext by means of the PRD (the percentage root-mean-square difference), its normalized version (PRDN) measures and NMSE (normalized mean square error).

The percentage root-mean-square difference (PRD) measure defined as (6):

$$PRD\% = 100\sqrt{\frac{\sum_{n=1}^{N}(x(n)-\tilde{x}(n))^2}{\sum_{n=1}^{N}x^2(n)}} \qquad (6)$$

is employed, where $x(n)$ is the original signal, $\tilde{x}(n)$ is the reconstructed signal, and N is the length of the window over which the PRD is calculated. The normalized version of PRD, PRDN, which does not depend on the signal mean value, $\bar{x}$, is defined as (7):

$$PRDN\% = 100\sqrt{\frac{\sum_{n=1}^{N}(x(n)-\tilde{x}(n))^2}{\sum_{n=1}^{N}(x(n)-\bar{x})^2}} \qquad (7)$$

The normalized mean square error (NMSE) measure defined as (8):

$$NMSE = \frac{1}{\sigma^2} * \frac{\sum_{n=1}^{N}(x(n)-\tilde{x}(n))^2}{N} = \frac{MSE}{\sigma^2} \qquad (8)$$

Where σ are the variance and MSE are mean square error measure.

Because our messages are text type, ie contain characters and numbers, we chosen to transform the messages in numerical signals based on the ASCII codes.

To use the identity matrix as decoding dictionary, the plaintext is necessary to be a sparse signal [18]. Because our messages had not this property, we have modified them by artificial insertion of zeros, thus obtaining sparse signals.

We used random matrix for encryption and for reconstruction we used two different algorithms, and namely,

- Basis pursuit algorithm (BP), known in the CS domain as the optimal algorithm in terms of errors [19-20] and
- Orthogonal matching pursuit algorithm (OMP) known in CS domain for its speed far superior to BP [21].

The orthogonal matching pursuit algorithm (OMP) is an iterative greedy algorithm. In this algorithm, at each step, the dictionary element which has the maximum correlation with the residual part of the signal is selected. The Basis Pursuit algorithm (BP) is a more sophisticated approach comparatively with OMP. In case of the BP algorithm, the initial sparse approximation problem is reduced to a linear programming problem.

Generically, the greedy algorithms (such OMP) have the disadvantage that there are not general guarantees of optimality. The basis pursuit algorithm, namely the convex relaxation algorithms, has the disadvantage of high computational complexity, translated into large computing time [22-26].

To synthesize ideas, we present the encryption and decryption necessary steps, namely:

- The message transformation into digital signal using extended ASCII code. This achieves a 1D digital signal.
- The segmentation of message or digital signal into segments of length 100.

- Transforming of the signals (signals with length 100) in sparse signals by inserting a predefined number of zeros. The position of the zeros is random from one segment to another.
- Encryption of sparse segments using a random matrix. Encryption is done by multiplying the signal sparse with a random matrix (Φ), resulting a lower dimension signal than initially sparse signal. The signal thus obtained is not sparse.
- Transmission of the message text is achieved by transmitting the encrypted signals (ciphertext) on an insecure line. It is important that random matrix (encryption matrix representing the secret key) is not sent with the ciphertext; it should be sent on a secure line. Another variant is use case when there is an agreement between the transmitter and receiver to generate random matrices in the same way, for example, using the same random number generator which is started from the same initial conditions.
- Decryption of the message will be achieved using a greedy algorithm (either orthogonal matching pursuit (OMP), or matching pursuit (MP), or greedy LS etc.) or convex relaxation algorithm (basis pursuit (BP)). For decryption, it is necessary to know the following: random matrix encryption Φ , the encrypted message (the ciphertext) Y, and the base for sparsity Ψ (in case of this paper, it is the identity matrix, due the fact that the message that was encrypted was a sparse signal).
- Because there is a decryption error which is very small, to return to the decrypted text, a decryption correction will be necessary. This correction consists in rounding of decrypted values to the nearest integer because the ASCII code is built from integers.

Figure 2 shows an example of plaintext and figure 3 presents a plot of the plaintext in ASCII format.

Anexa nr.1
DECLARATIE
privind veniturile realizate
Agentia Nationala de Administrare
200
Fiscala
din România
Anul Se completeaza cu X în cazul declaratiilor rectificative
A. DATE PRIVIND ACTIVITATEA DESFASURATA Cod CAEN cote forfetare de cheltuieli norma de venit Nr. Data 7. Data începerii activitatii 4. Obiectul principal de activitate 5. Sediul/Datele de identificare a bunului pentru care se cedeaza folosinta 8. Data încetarii activitatii asociere fara personalitate juridica entitati supuse regimului transparentei fiscale individual 6. Documentul de autorizare/Contractul de asociere/Închiriere/Arendare 3. Forma de organizare: 2. Determinarea venitului net: comerciale profesii libere drepturi de proprietate intelectuala cedarea folosintei bunurilor operatiuni de vânzare-cumparare de valuta la termen, pe baza de contract transferul titlurilor de valoare, altele decât partile sociale si valorile mobiliare în cazul societatilor închise activitati agricole 1. Categoria de venit Venituri: cedarea folosintei bunurilor calificata în categoria venituri din activitati independente sistem real modificarea modalitatii/formei de exercitare a activitatii

Fig. 2. The plaintext

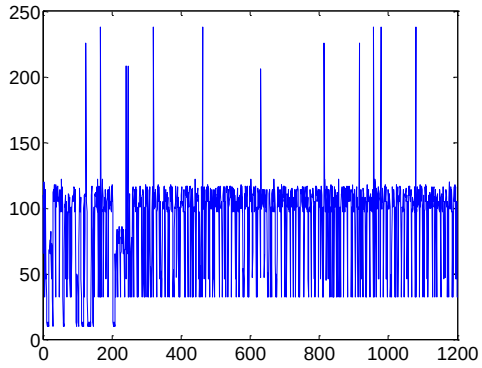

Fig. 3. The plaintext in ASCII format

We have chosen to split the signal into segments of length 100 and to insert a number of 800 by zeros for each ASCII codes segment. This means that each plaintext sequence with length 100 was transformed into a sequence with length 900. Figure 4 shows the plot of sparse plaintext.

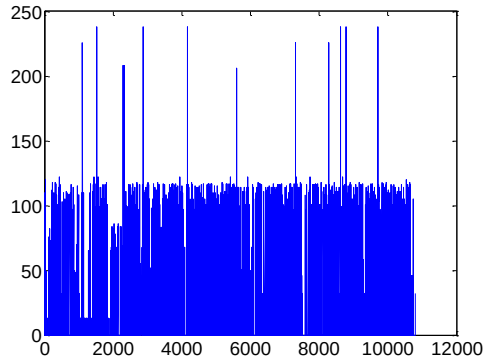

Fig. 4. The sparse plaintext

We used for encryption a random matrix of size 500x900. This random matrix represents the secret key. Figure 5 show the ciphertext obtained a random matrix for encryption. Note that the ciphertext contains positive and negative numbers and it has a different length than the plaintext.

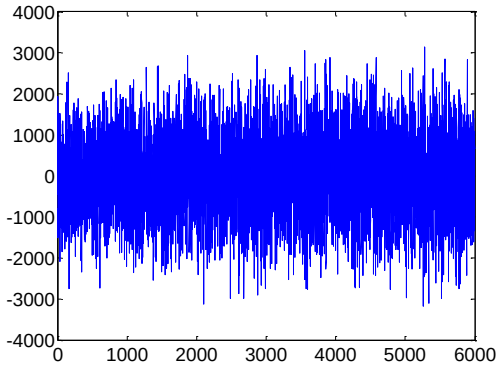

Fig. 5. The ciphertext

To decode the ciphertext we tested two known algorithms from compressed sensing domain, namely, orthogonal

matching pursuit algorithm (OMP) and basis pursuit algorithm (BP).

OMP is an iterative greedy algorithm and selects at each step the column of Φ matrix which has the maximum correlation with the current residuals. A set of iteratively selected columns is built. The residuals are iteratively updated by projecting the observation y onto the subspace spanned by the previously selected columns. This algorithm has simpler and faster implementation toward similar methods.

The Basis Pursuit (BP) algorithm consists in finding a least L1 norm solution of the underdetermined linear system $y = \Phi x = \Phi \Psi \gamma$.

The both methods can be guaranteed to have bounded approximation solution of sparse coefficients estimation for the condition that the L0 norm of sparse coefficients is smaller than a constant decided by the dictionary [1].

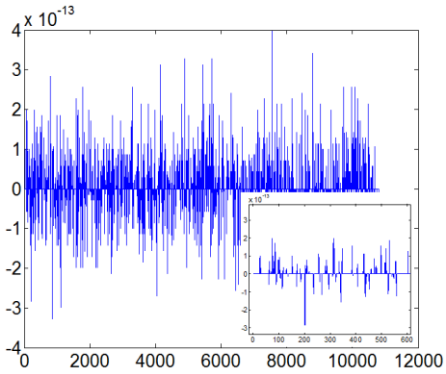

Fig. 6. Error for decoding with OMP, before decoding correction. In the bottom right corner there is a zoom for the first 600 samples

Figure 6 and figure 7 show errors for decoding with OMP, respectively BP algorithms.

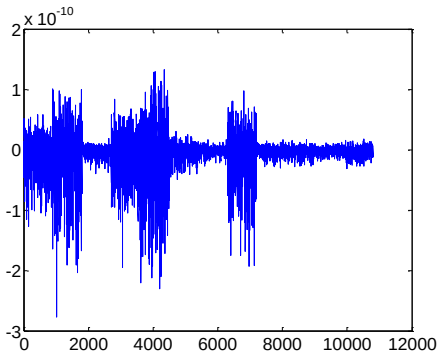

Fig. 7. Error for decoding with BP, before decoding correction

For the OMP based decoding, where the original signal (plaintext) was sparse (had null values), null values were obtained after decoding. In case of BP decoding, the algorithm approximates all values and it failed to return null values for the null values from plaintext, but it returned values very close to zero.

Because in the case of typical tax forms often it is required to encrypt only registration data and because the decryption time is higher for the completed form (data + template), we tested the proposed algorithm for encrypting data alone. In Figure 8 is an example of data belonging to the form shown in Figure 2.

12656
03.08.2010
Minerit
Str. Minei, Nr. 23, Hunedoara
02.08.2000
contract nr. 05/09.10.1999
SRL

Fig. 8. The plaintext with registration data from tax form

For a signal of dimension m with assumed sparsity s<<m, and a dictionary of N>>m atoms, computational costs for pursuits using general and fast dictionaries are:

$$complexity \quad for \quad OMP = smN(sN \log N + s^3))$$

where m stands for measurements, s stands for sparsity.

The popular basis pursuit algorithm (BP) has computational complexity

$$complexity \quad for \quad BP = O(N^3)$$

Alternatives to BP (e.g., greedy matching pursuit) also have computational complexities that depend on N.

Table 1 presents average results for 10 text messages and for 10 datasets from tax forms. The time for BP algorithm is much higher than for OMP algorithm and the errors are smaller.

TABLE I. AVERAGE RESULTS

Decoding algorithm	Time (seconds)	Error (PRD, PRDN, NMSE)
average results for 10 text messages, each with 1200 char		
Basis pursuit algorithm (BP)	867.40	PRD = 7.7521e-011 PRDN = 8.1579e-011 NMSE = 7.1476e-028
Orthogonal matching pursuit algorithm (OMP)	2.61	PRD = 1.0139e-013 PRDN = 1.0670e-013 NMSE = 1.2227e-033
average results for 10 registration data text messages, each with 103 char		
Basis pursuit algorithm (BP)	42.27	PRD = 3.9552e-011 PRDN = 4.1392e-011 NMSE = 3.2769e-028
Orthogonal matching pursuit algorithm (OMP)	0.09	PRD = 1.0979e-013 PRDN = 1.1489e-013 NMSE = 2.5248e-033

It should be noted that the OMP does not guarantee the convergence of the algorithm and for a smaller number of measurements; the results can be much worse for OMP comparatively with BP. Results depend on the number of measurements and on used decoding algorithm [24-26].

IV. CONCLUSIONS

In this paper, the perfect secrecy via compressed sensing was studied and discussed. We presented an analysis with practical results for tax forms as plaintexts. For decoding we used BP and OMP algorithms, and we presented a comparative analysis. The time for BP algorithm is much higher than for

OMP algorithm and the errors are smaller and should be noted that the OMP does not guarantee the convergence of the algorithm. According to average results from Table 1, it is more advantageous, for tax forms (or other templates that show no interest for encryption) to encrypt only the recorded data. The time required for decoding is significantly lower than the decryption for the entire form.

ACKNOWLEDGMENT

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS – UEFISCDI, project number PN-II-RU-TE-2014-4-0832 "Medical signal processing methods based on compressed sensing; applications and their implementation."

REFERENCES

[1] D. Donoho, "Compressed sensing", IEEE Transactions on Information Theory, vol. 52(4), pp. 1289–1306, 2006

[2] E. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information" IEEE Transactions on Information Theory, vol. 52(2), pp. 489–509, 2006.

[3] M. Elad, "Optimized Projections for Compressed Sensing", IEEE Transactions on Signal Processing, Vol. 52, 2007

[4] Shuhui Bu, Zhenbao Liu, Tsuyoshi Shiina, Kazuhiko Fukutani, *Matrix Compression and Compressed Sensing Reconstruction for Photoacoustic Tomography*, Elektronika ir elektrotechnika, Vol 18, No 9 (2012)

[5] C. E. Shannon, "Communication theory of secrecy systems" Bell System Technical Journal, vol. 28(4), pp. 656–715, October 1949.

[6] E. Candes, J. Romberg, "Sparsity and incoherence in compressive sampling" Inverse Problems, Vol. 23, pp. 969–985, 2007.

[7] D. Takhar, J. N. Laska, M. B. Wakin, M. F. Duarte, D. B. S. Sarvotham, K. F.Kelly, and R. G. Baraniuk, "A new camera architecture based on optical-domain compression" in Proc. IST/SPIE Symposium on Electronic Imaging: Computational Imaging, vol. 6065, 2006, pp. 129-132

[8] Y. Rachlin, D. Baron, "The Secrecy of Compressed Sensing Measurements", 46th Annual Allerton Conference on Communication, Control, and Computing, 2008

[9] A. Orsdemir, H. Oktay Altun, G. Sharma, Mark F. Bocko,"On the Security and Robustness of Encryption via Compressed Sensing" Military Communications Conference, 2008, Milcom 2008, IEEE pp.1-7

[10] G. Zhang, S. Jiao, X. Xu, "Application of Compressed Sensing for Secure Image Coding", Lecture Notes in Computer Science Volume 6221, 2010, pp 220-224.

[11] Gary C. Kessler, An Overview of Cryptography, 2015, http://www.garykessler.net/library/crypto.html

[12] Sattar B. Sadkhan Al Maliky and Nidaa A. Abbas, Multidisciplinary Perspectives in Cryptology and Information Security, IGI Global, 2014

[13] W. Stallings, Cryptography and Network Security (4th Edition), pp. 30, Prentice Hall, 2005

[14] V. Preoteasa, Cryptography and Network Security, Lecture 2: Classical Encryption Techniques, Spring 2008, Abo Akademi University

[15] D.R. Stinson, Cryptography: Theory and Practice, 2nd edition, Chapman & Hall/CRC, 2002

[16] W. Diffie, M. E. Hellman, "New directions in cryptography" IEEE Transactions on Inform. Theory, vol. IT-22, no. 6, pp. 644–654, 1976.

[17] U. Maurer, S. Wolf, "Information-theoretic key agreement: From weak to strong secrecy for free" Advances in Cryptology—EUROCRYPT, Lecture Notes in Computer Science, 2000.

[18] J. Bowley, L. Rebollo - Neira, "Sparsity and "something else": an approach to encrypted image folding", IEEE signal processing letters, 18 (3), pp. 189-192., 2011

[19] D. Donoho, Y. Tsaig, "Fast solution of L1-norm minimization problems when the solution may be sparse," Stanford University Department of Statistics Technical Report, 2006.

[20] D. Donoho, "For most large underdetermined systems of linear equations, the minimal L1 norm solution is also the sparsest solution" Communications on Pure and Applied Mathematics, Vol. 59, pp. 797–829, June 2006.

[21] J. Tropp, A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit" IEEE Trans. on Information Theory, Vol. 53, No. 12, pp. 4655–4666, December 2007.

[22] E. Candes, T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies?", IEEE Transactions on Information Theory, Vol. 52, No. 12, pp. 5406–5425, December 2006.

[23] M. J. Wainwright, "Sharp thresholds for noisy and high-dimensional recovery of sparsity using L1-constrained quadratic programming (Lasso)", IEEE Transactions on Information Theory, 2009.

[24] T.T. Cai, L. Wang, "Orthogonal Matching Pursuit for Sparse Signal Recovery with Noise", IEEE Transactions on Information Theory, vol. 57, 7, 4680–4688, 2011.

[25] G. Davis, S. Mallat, M. Avellaneda, "Greedy adaptive approximation", J. Constr. Approx., 13:57-98, 1997.

[26] J.A. Tropp, "Greed is good: Algorithmic results for sparse approximation", IEEE Transactions on Information Theory, 50, pp. 2231–2242, 2004.

[27] M. N. Chavhan, S.O.Rajankar, "Study the Effects of Encryption on Compressive Sensed Data", International Journal of Engineering and Advanced Technology, Volume 2, Issue 5, pp. 179 – 182, 2013