A Novel Method to Protect Content of Microsoft Word Document Using Cryptography and Steganography

Mohamed Ahmed Mohamed, Obay G. Altrafi, Mohamed O. Ismail, and Mawada O. Elobied

***Abstract*—Microsoft Word is the most famous and popular word processor in the world nowadays and the last versions of it depend on eXtended Markup Language (XML) by using Open Office XML format (OOXML). Usually MS Word document contain sensitive data, this document may be sending over internet, via email or even stored in cloud. This makes the data in this document vulnerable to many security threats. This paper introduce novel method to protect the data inside MS-word document using cryptography and steganography techniques, via propose two algorithms, one for hiding user's selected content from the document into an zero dimension image, and the other for retrieving the original plain text. The implemented method gave a good result in hiding the content of the file regard less the type of the content.**

***Index Terms*—MS-Word document, cryptography, steganography, XML fileK, zero dimension image, MS-Word add-on.**

I. Introduction

Microsoft Word is the most famous and popular word processor in the world nowadays and the last versions of it depend on eXtended Markup Language (XML) by using Open Office XML format (OOXML).

Cryptography and steganography are techniques used to grantee confidentiality and concealment of several kinds of data such as image, video, and text.

Usually MS Word document contain sensitive data, this document may be sending over internet, via email or even stored in cloud. This makes the data in this document vulnerable to many security threats.

In this paper, we tried to introduce a new method to protect these sensitive data and the main goal of our new method to protect only contents of document that the user wants to protect from document. By implementing Add-in compatible with MS-Word Application to provide User Interface (IU) and also doing main process like encryption/decryption and hide/show content.

The structure of this paper started with brief overview of Microsoft Word, cryptography and steganography in Section II. We addressed related work in Section III. The proposed work and methodology to implement this method discussed in Section IV. The experimental result and discussion was shown in Section V. Finally, we concluded our work in Section VI.

Manuscript received March 18, 2014; revised June 3, 2014. This work was supported in part by University of Khartoum.

Mohamed Ahmed Mohamed is with the ICT Department, The National Assembly, Sudan (e-mail: Mohamed.alfaki@hotmail.com).

II. Overview Microsoft Office Word an Cryptography and Steganography

A. Microsoft Office Word

Microsoft Word developed by Microsoft in 1983 and since then many versions been released. The name word come in 1983 before that its name was Multi-Tool Word which works only on the Xenix systems but the later versions worked in many platforms and operating systems include DOS, Apple Macintosh, Unix, OS/2, Atari ST, SCO UNIX , and Windows.

Part of the success of Microsoft Word is the popularity of Windows operating system itself because throw out the years many versions of Microsoft Office suites has introduced, nonetheless the first version of Office suites just contained Microsoft Word, Microsoft Excel, and Microsoft PowerPoint. Furthermore, it considered the main program in Office.

Microsoft Word uses special kind documents known as MS-word documents to store its information. All the versions of MS-word until 2003 used MS-word document that end by the excitation ".doc" which became a de facto standard of document file formats for Microsoft Office users. This Binary File Format implements Object Linking and Embedding (OLE) structured storage to manage the structure of their file format. After that Microsoft, introduced new version (Microsoft Office 2007) using XML format as structure for document including OLE compress by ZIP algorithm as one file that end by ".docx" excitation [1].

Microsoft Word application has many features that added in newer versions, which add much functionality. *Document Inspector* is one of these features, which added in order to guarantee a significantly high level of privacy and security, which makes it possible to quickly identify and remove any sensitive, hidden and personal information [2].

In addition, Microsoft Word application allows the possibility of adds a new functionality for the main features called *Add-in*. it can develop by using Visual Studio to customize MS-Word application and add the specific features we need for our business processes or customize UI of application. For example, we can turn Word into a contract generator that assembles contracts out of pre-existing parts that can be made editable or not editable [3].

Content controls provide a UI that is optimized for both user input and print. When you add a content control to a document, the control is identified by a border, a title, and temporary text that can provide instructions to the user. The border and the title of the control do not appear in printed versions of the document. There are nine different types of content controls that you can add to documents. Most of the

content controls have a corresponding type in the (Microsoft Office Tools Word) namespace. You can also use a generic (Content Control), which can represent any of the available content controls. One of this the Rich Text control one of this types, contains text or other items, such as tables, pictures, or other content controls [4].

B. Cryptography and Steganography

Encryption is a process of coding information which could either be a text, file or mail message into cipher text to make it unreadable without a decoding key in order to prevent anyone except the intended recipient from reading that data. Decryption is the reverse process of converting encoded data to its original un-encoded form, plaintext. A key in cryptography is a long sequence of bits used by encryption / decryption algorithms to encrypt plain text or to decrypt cipher text. A given encryption algorithm takes the original message with a key, and alters the original message mathematically based on the key's bits to create a new encrypted message. Likewise, a decryption algorithm takes an encrypted message and restores it to its original form using one or more keys. There are many encryption algorithms used nowadays to encrypt and decrypt our file, text, image, audio and video. There are some algorithms like Data Encryption Standard (DES), International Data Encryption Algorithm (IDEA), Blowfish, Triple DES (TDES), Twofish, RSA, Diffie-Hellman, Elliptic Curve Cryptography (ECC), Pretty Good Privacy (PGP), Public key infrastructure (PKI) [5].

AES (Advanced Encryption Standard) one of these algorithms. The AES algorithm is a symmetric block cipher that can encrypt and decrypt information. The AES algorithm is capable of using cryptographic keys of 128, 192, and 256 bits to encrypt and decrypt data in blocks of 128 bits [6].

Many application need to use cryptography algorithms must define the key it will use and almost of these application need strong key and used what it call hash function to make strong key. Hash Function operate on messages of almost arbitrary length and output a fixed size value. Cryptographic hash functions should satisfy many security properties, such as the impossibility from a given hash to recover an associated message. However, the main security requirement for a hash function is its collision resistance [7].

In practice, building a cryptographic function with an input of variable size is not a simple task. For this reason, most hash functions are based on an iterated construction that makes use of a so-called compression function, whose inputs have fixed sizes. Examples of such a construction are Snefru, MD4, MD5 or SHA [8].

Other method to protect data is steganography, which the art and science of inconspicuously hiding data within data, so main goal of steganography is to hide information well enough to prevent the unintended recipients from suspecting the stenographic medium containing any hidden data. There are many stenography techniques, which use many mediums to hide data inside it; these mediums are called the cover mediums. Furthermore the methods that used to hide data into cover medium is called embedding techniques, the two most common methods used for hiding information inside a Document is to embed information inside a document we can simply alter some of its characteristics .i.e. either the text formatting or characteristics of the characters [9].

III. Related Work

In 2011, Prof. Dr. Abdul Monem S. Rahma *et al*., this paper proposed and implemented method to hide data in unused block of binary file format of Microsoft Compound Document File Format (MCDFF), they gave positive results and using Track Changes tool does not effect on hidden data and no problem was detect on hidden data at stego-document mailing or copying [10].

In 2012, Jassam *et al*., proposed an effective method for hiding data in Microsoft word documents. This method contain two algorithms: first algorithm used to hide secret message in HMTL color property after convert document to HTML format, and second algorithm used to retrieve original message from the cover document. The web scripting language, personal homepage (PHP) was use to implement their method; the experimental result of study show an effective result for English and Arabic languages [11].

In 2012, Abikoye Oluwakemi C. *et al*., a system that combined the techniques of cryptography and steganography to provide efficient method of hiding data from any unauthorized users was present. A data hiding system that is based on audio steganography and the Least Significant Bit algorithm was employed to encode the message inside the audio file [12].

In 2013, Wesam Bhaya *et al.*, proposes a novel text steganography method that takes into account the Font Types. This new method depends on the Similarity of English Font Types and it works by replace font by more similar fonts. The secret message was encoded and embedded as similar fonts in capital Letters of cover document. This new method called Similarity of English Font Types (SEFT) technique [13].

IV. Proposed Work and Methodology

A. Proposed Work

Our proposed method is using an encryption technique to encrypt selected content (text, images, tables …) by the user inside MS-Word documents and after that hide these encrypted contents using Zero Dimension Image steganography technique.

This method consist of two algorithms, one is to hide user's selected content in the document and other to retrieve original plain text.

When the user selects some contents to be hidden inside a document, the user is obligated to enter password that used in encryption process. The password will be hashed using MD5 hash function, to generate fixed length password of 128-bit size to be used in encryption algorithm.

The user select content in word document that need to be protected, getting XML that represents the selected section, and after that this XML is encrypted by AES-128 encryption algorithm and key for this process is the 128 bit hash function result which entered by the user. The cipher text is hidden into Zero dimension image, and this image is replace in

position of the select section into document as shown in Fig 1. Second algorithm, to reverse processes in algorithm one to retrieve original content and Fig. 2 explain these process.

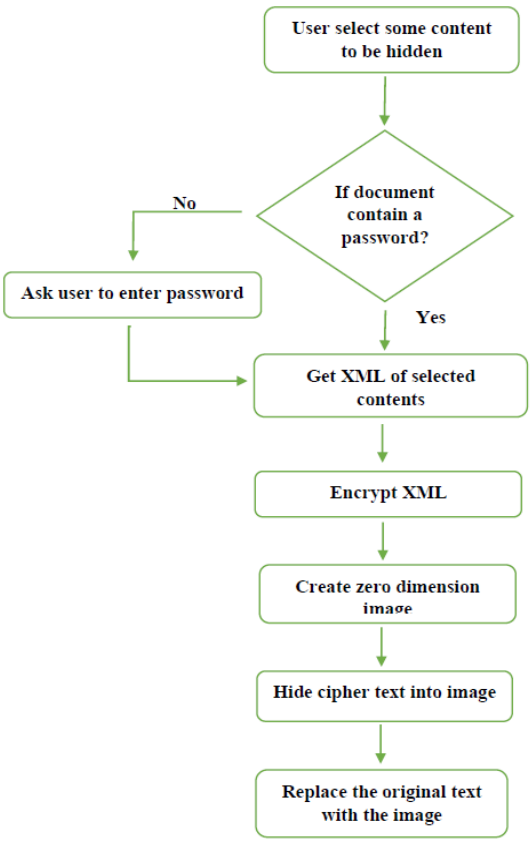

Fig. 1. The proposed algorithm for hiding the contents.

B. Methodology

We divided the system to two of processes, Encrypt/Hide process and Decrypt/Show process.

1) Encrypt/hide process

Encrypt: XML content of the selected section encrypted via AES encryption algorithm with key length of 128 bit. The key is acquired from the user and then hashed using MD5 hash function, which produce 128 bit hashed value.

Embed into image: the cipher text that produced by the encryption process embed into an image using binary file steganography technique in which the hidden data is added by making changes to the binary code that does not affect the execution of the file [9].

Hide process: the cover media, which is an image that contain the embedded data, is resized to zero dimension image and even its color is changed to match the document background so it is impossible to be noticed by the naked eye.

2) Decrypt/show process

Extract: the document searched for each zero dimension image, the cipher text is extracted using same the binary file steganography technique we used to hide the data in the first place.

Decrypt: cipher text is decrypted by AES algorithm, by using the hashed of the key that been entered by the user.

Show: the plan text that result from decryption process, replaces the zero dimensions image.

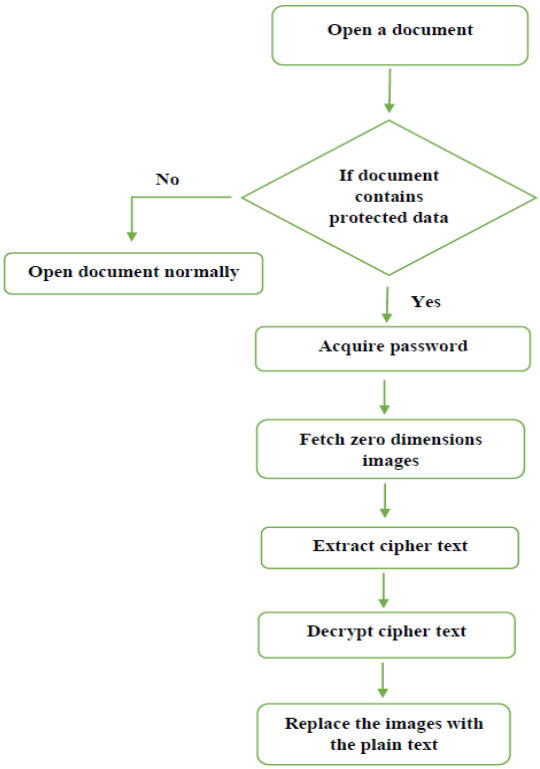

Fig. 2. The proposed algorithm for extracting the contents.

V. Experimental Result and Discussion

To prove the proposed method a system was developed using C#.NET language on .NET framework 3.5 and Microsoft Visual Studio Tools for Office (VSTO). This system is an MS word add-on, which an application level add-on that appears in the ribbon as a tab which contains user interface (UI) for our program as shown in Fig. 3(1), the UI consist of two control groups, the first group is *encryption section* with button called *encrypt selection* which encrypt the selected content as shown in Fig. 3(2), and *password at open* check box, which give the user a choice of ask the user who open the document, of password immediately after he opens the document or after he click show button in decryption group.

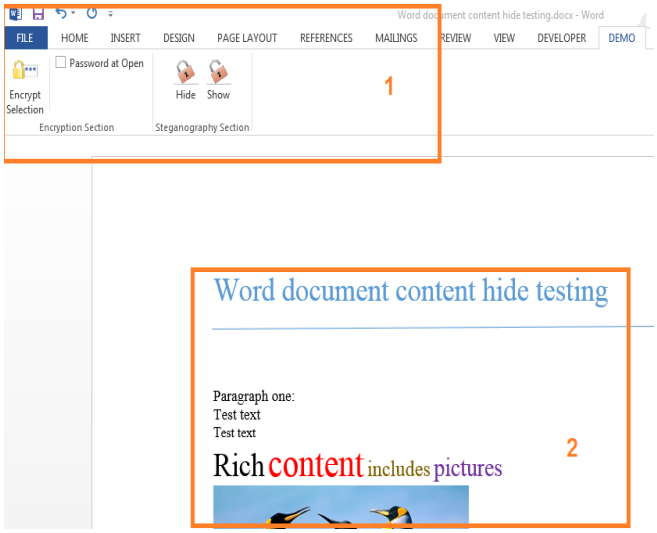

Fig. 3. Illustration of (1) ribbon tab and (2) document contents.

The second group is steganography section which, consist

of two buttons show button and hide button and as there name refer to their functionality show button is used to show the hidden content after entering password if the user doesn't enter it at the opening of the file. The hide button is used to hide the content after they firstly encrypted.

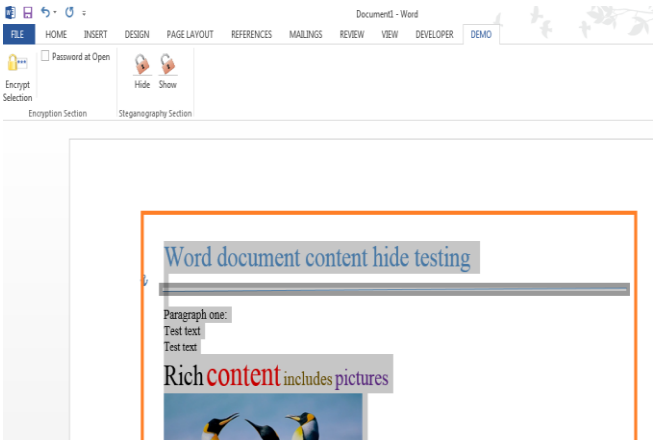

Fig. 4. Illustration selection process.

A. Encrypt/Hide Process

As shown in Fig. 4 when the user select the content and click encryption as in Fig. 5(2) a password window will appear as illustrated in Fig. 5(1) and the user has to enter password and it will be hashed to give the user more flexibility and freedom in entering any password length, also the hashing of the password improve the security because we don't use the password itself.

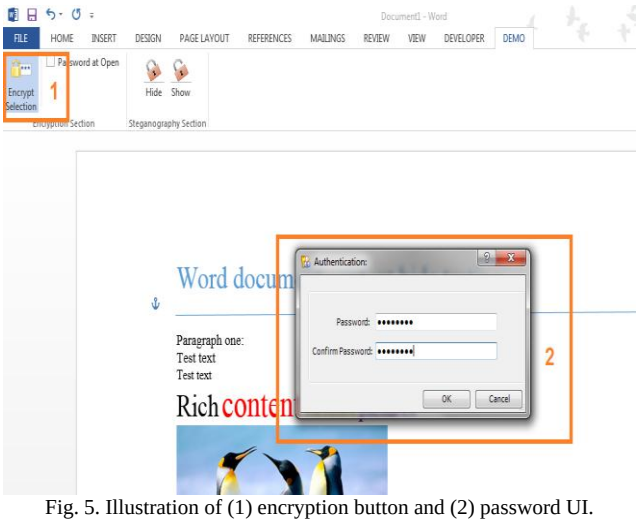

Fig. 5. Illustration of (1) encryption button and (2) password UI.

When the user click OK after entering the password Fig. 5(2) the selected text will be contained inside a rich text control box as showing in Fig. 6(2) to give the user the ability of editing and modification selected content inside the box before of the hiding process, after that when the user click hide button in the steganography section as shown in Fig. 7(1), the content of the rich text control box will be hidden into an zero dimension image and replace the original content as shown in Fig. 7(2). The zero dimension image colour is white and it not never be noticed by the naked eye. The image also will be contained inside rich text box in order to be protected from deletion accidently for example when the user select all the content and delete it.

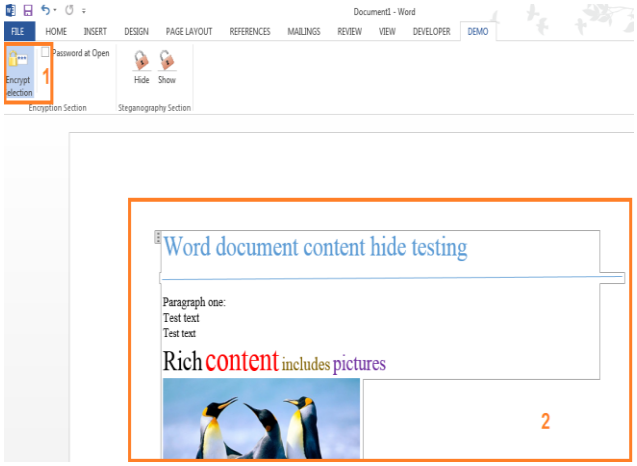

Fig. 6. Illustration of (1) encryption button and (2) selected content bordered by rich text box control.

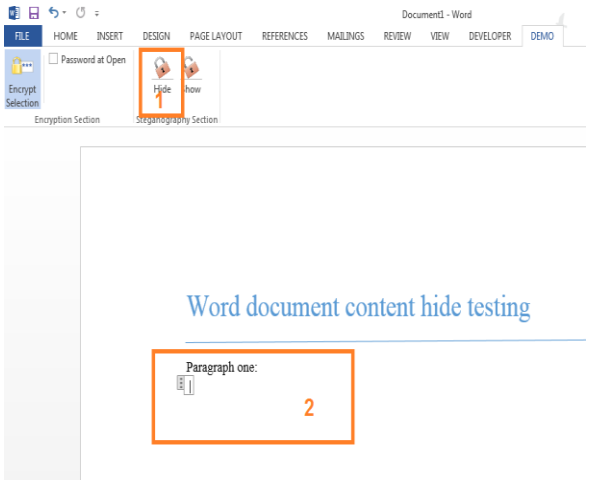

Fig. 7. Illustration of (1) hide button and (2) content hide process.

B. Decrypt/Show Process

When the user click show button in the steganography section the encrypted content will be decrypted with hash of the user entered password and each image will be replaced with its original decrypted content.

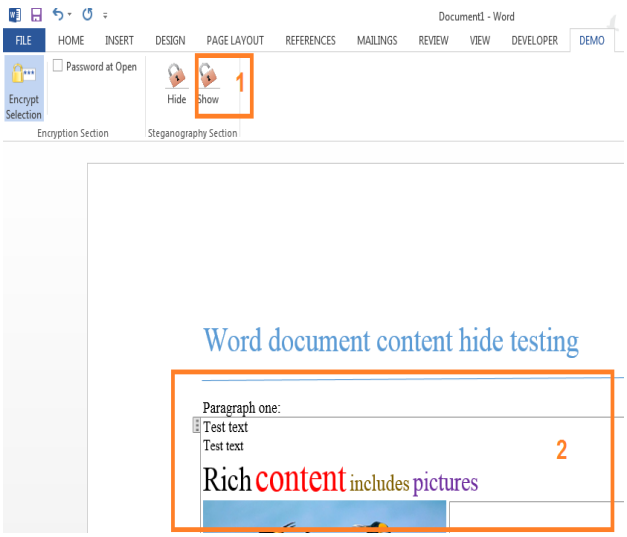

Fig. 8. Illustration of (1) show button and (2) content show process.

During the implementation some issues is raised with the box that contain the zero dimension image like the issue of the deletion that will compromise the integrity of the document, and we solve this problem.

C. Discussion

the implemented method give good result in hiding the content of the file regard less the type of the content (image, rich text...) and kipping the format of the selected content. So when the retrieve it he does not worry about the format mixing, the test was done on Arabic and English text.

When the document that contains hidden data checked using MS word inspector, it gives positive result and the inspector could not detect the hidden data.

VI. Conclusion

In this paper, a system that combined the techniques of cryptography and steganography to provide efficient method of hiding data from any unauthorized users was presented. MS-word document was used as medium to hide the content using zero dimension image technique. The method prove efficiency in hiding any type of content, the size of the file is increasing by relatively big amount due to ciphering so we recommend using compression technique to compress the cipher text before hiding it.

References

[1] R. Shah and J. Kesan, "Interoperability challenges for open standards: ODF and OOXML as examples," in *Proc. the 10th Annual International Conference on Digital Government Research: Social Networks: Making Connections between Citizens, Data and Government* , 2009, pp. 56-62.

[2] A. Castiglione, B. D'Alessio, A. D. Santis, and F. Palmieri, "Hiding information into OOXML documents: New steganographic perspectives," *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, vol. 2, no. 4, p. 25, 2011.

[3] Microsoft. (2013). Microsoft Web page. [Online]. Available: http://msdn.microsoft.com/en-us/library/office/aa189710%28v=office.10%29.aspx

[4] T. D. M. Sunderland, "Techniques to create structured document templates using enhanced content controls," United States Patent US 2012/0254730 A1, 4 October, 2012.

[5] A. Kakkar, M. L. Singh, and P. Bansal, "Comparison of various encryption algorithms and techniques for secured data communication in multinode network," *International Journal of Engineering and Technology*, vol. 2, no. 1, p. 6, January 2012.

[6] *Advanced Encryption Standard (AES)*, NSIT Standard 197. 11-2001.

[7] *Secure Hash Standard*, NSIT Standard 180-2. 8-2002.

[8] A. Joux, *Multicollisions in Iterated Hash Functions Application to Cascaded Constructions*, Paris, 2004.

[9] A. J. S. Channalli, "Steganography an art of hiding data," *International Journal on Computer Science and Engineering*, vol. 1, no. 3, pp. 137-141, 2009.

[10] A. M. S. Rahma, B. AbdulWahab, and A. Y. Al-Noori, "Proposed steganographic method for data hiding in Microsoft word documents structure," *Al-Mansour Journal*, no. 15, pp. 1-29, 2011.

[11] J. Sarsoh, K. Hashem, and H. Hendi, "An effective method for hidding data in Microsoft Word documents," *Global Journal of Computer Science and Technology Network, Web & Security*, vol. 12, no. 12, pp. 39-42, 2012.

[12] A. Oluwakemi, A. Kayode, and O. Ayotunde., "Efficient data hiding system using cryptography and steganography," *International Journal of Applied Information Systems (IJAIS)*, vol. 4, no. 11, pp. 6-11, December 2012.

[13] W. Bhaya, A. Rahma, and D. Al-Nasrawi, "Text steganography based on font type in ms-word documents," *Journal of Computer Science* , vol. 9, no. 7, pp. 898-904, 2013.

Mohamed Ahmed Mohamed was born in Khartoum in 1987. He received the BSc. degree in computer systems and networks from University of Sudan in 2009, and his MSc degree in computer science from University of Khartoum, Khartoum, Sudan.

He joined The National Assembly Sudan since 2011 as a computer engineer, before that he worked as a software engineer in Micronet Integrated Solutions Co Ltd.

His main areas of research interest are information security, databases, web technologies and software engineering.

Obay Gsemalseed Ahmed Altrafi was born in Aljazeera, Sudan on 30, May 1989. He received the bachelor's degree in computer systems and networks from Sudan University of Science and Technology in 2011. He acquired his master's degree in computer science field of information security from University of Khartoum in 2013, now he is a Ph.D. candidate in computer science in King Fahd University of Petroleum and Minerals (KFUPM), KSA, 2014. His research interest is data protection, big data issues, cloud computing, mobile computing, ubiquitous computing, wireless network, operating system etc.

.

Mohamed Osman Ismail was born in Sudan on March 02, 1988. He received the BSc. degree in computer systems and networks from Sudan University of Science and Technology, Faculty of Computer Science, Khartoum, Sudan, in 2009.

In 2014, he received the MSc. degree in computer science in the subject of information and communication systems security from University of Khartoum, Faculty of Mathematical Sciences, Khartoum, Sudan.

He is currently a software developer in the Department of Software Development in Micronet Integrated Solution Company since 2009.

His research interests include information security, network security and the problem of storage big data in database.

.

Mawada Osman Elobied was born in 1988. She received the bachelor's degree in computer science from University of Khartoum, Faculty of Mathematical Sciences in 2009. She acquired her master's degree in computer science field of information security from University of Khartoum in 2013.

Now she is looking forward to improve her skill in software developing. Her research interest is business intelligent (BI) by using smart phone and the toll that used in securing the sensitive information or data.

.