# An Efficient and Secure Approach for Data Lineage Detection in Malicious Environments

Sincija C[1], Dr. Dhanabal S[2], Silja Varghese[3]

[1]*M.tech., Student,* [2]*Associate Professor,* [3]*Assistant Professor*

*Computer Science and Engineering, Nehru College of Engineering and Research Centre*

*Thrissur, Kerala, India*

***Abstract— In the digital era leakage of confidential or sensitive data has become one of the most security threats to the organizations. The data leakage can be either intentional or unintentional. The increased use of social network services plays a role in data leakage. By the use of those services our personal information can be made available to the public very easily. And also the privacy preserving mechanisms are limitted to those services. Some of the smartphone services also can unknowingly shares our personal data to the untrust third party applications. Data lineage detection by the use of watermarking techniques and fake data addition, has already suggested and has been implemented by some of the organizations. Most of them was ad-hoc in nature and there was no formal model available. The proposed system called LIME provides a framework for data lineage detection. It provides the exact security guarantees for data flow between different entities using steganography. The application of the system is for the identification of the guilty entity and also to detect the non-repudiation and honesty assumptions in data transfer. It also provides protection to the data by blocking the data leak.***

***Keywords: Lineage, LIME, Entities, Owner, Consumer, Auditor.***

## I. INTRODUCTION

The confidential information leakage through unintentional exposures and malicious external entities, is one of the most serious security threats to organizations in the digital world. A large amount of digital data can be copied at low cost. Those data can be spread through the internet in a very short span of time. As there was only a limited accountability mechanisms currently available, the risk of getting caught for data leakage was very much less. Due to these reasons, the problem of data leakage has become an important topic for the research today.

The leakage of the confidential data has affected the security and privacy of the organizations widely. It has become a concern to the individuals too. The increased rise of social networks and smartphones plays a role in data leakage and has made the situation some more worse. These environments makes the individuals to disclose their personal information to their service providers, mainly to the third party applications for getting some of the services which are of free cost. Due to the absence of proper regulations and accountability mechanisms, many of those applications will share individual’s identifying information with other advertising and internet tracking companies. The sensitive data can be protected by the use of access control mechanisms, where access to data can be limited. Eventhough a malicious unauthorized user can publish those sensitive data at their concern. Information security primitives like encryption offers protection. The protection will be there only as long as the information is encrypted. But when the recipient decrypts that message, the security gets lost. The recipient can be able to pubish those document if he wish. Thus it seems to be very difficult to prevent the data leakage effectively.

The third party applications of the widely used online social networking sites can leak sensitive private information about the users or even their friends to advertising companies. Due to the improper outsourcing of the outsourcing company can also leads to the personal information or data leakage. Most of the data leakage scenarios are associated to an absence of accountability mechanisms during data transfers. That means, if entities know that they can be held accountable for leakage of data, they will demonstrate a better commitment towards its required protection.

The identification of the leaker can be made possible by forensic techniques in some cases, but they are expensive and do not always generate the desired results. Therefore, this research points out the need for a general accountability mechanism in data transfers. This accountability can be associated with detecting a transmission history of data across multiple entities starting from its origin. This method is known as data lineage, data provenance or source tracing. The data lineage methodology, in the form of robust watermarking [1] techniques and also adding fake data, has been already suggested and deployed by some of the organizations. But most of the efforts have been developed for a particular function and there was also no formal model available. And also, most of these approaches only allows for the identification of the leaker in a non-provable manner, which was not sufficient in many of the cases.

## II. SYSTEM MODEL

To address the general case of data leakage in data transfer settings proposed a simple model named LIME (Lineage In Malicious Environment). LIME is a generic data lineage framework for data flow across multiple entities. The entities in data flows assume one of two roles, it can be an owner or a consumer. In this research introduced an additional role in the form of auditor, whose task is to determine a guilty party for any data leakage, and also to define the exact properties for communication between these roles. The auditor is only required when a leakage occurs. The auditor then reconstructs the data lineage. The identification of the guilty entity or the data leaker can be made possible by the document owner from the lineage of leaked data provided by the auditor.

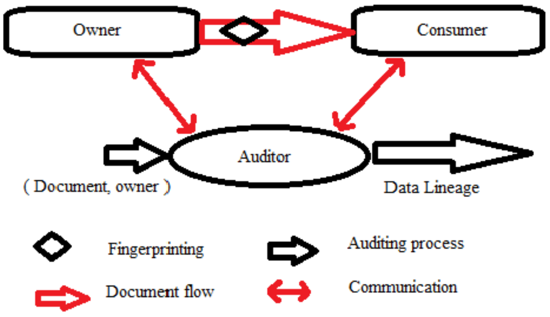


Fig. 1 LIME framework (Michael Backes et al., 2016, p. 3)

The LIME framework [1] is shown in the above fig.1 which shows data transfer between different entities. The entities involved in data flow includes the Owner, Consumer and Auditor. The owner is responsible for the management of documents in which it includes the creation, editing and modification of the documents. The owner has to embed the identifiers into the document so that those embedded information can uniquely identify the consumer or the recipient. Hence the owner first fingerprints their document and sent it to the consumer. At the same time the fingerprinted document having the identifier of the owner is sent to the auditor. Hence from that information the auditor can detect the data leakage and can inform about the details regarding the leaker to the owner. The consumer is the person who is in search for the required documents for their particular function. The consumer receives the documents and can carry out some tasks using them. The owner after fingerprinting their document directly transfers documents to consumers as shown in fig.1. The consumer can use the document provided by the owner to do their particular functions. But they cannot upload the document. Whenever a consumer tried to upload owner's document, the auditor will be invoked. Then the details regarding the leaker will be sent to the owner. The auditor is only invoked when a leakage occurs by the owner. Then the owner will provide the leaked document to the auditor. Then auditor will identify the leaker by using the embedded information done by owner at the time of fingerprinting. Then the auditor generates the data lineage which is the list of consumers other than document owner who tried to upload the document. Then the details about the unauthorised user will be given to the owner and performs all steps to identify the leaker.

## III. PREVIOUS WORK

The fast development of user friendly online social networks and smart mobile devices, sharing of photos becomes very easy and it has become a part of daily life. A large number of photos are being uploaded and shared at every fraction of second. Photos shared can be accessed and commented easily by the public. Most of these services lack a sound solution for protecting user's privacy. Social networks offers privacy protection solutions only in response to the public demand. They can provide only a limited degree of control and protection.

In [1] Michael Backes suggested an approach for data lineage detection by the method of robust watermarking using cox algorithm. As watermarked documents are easily to detect, to provide robustness watermarking was done on the core part of the document in the system. So that if the watermark is removed the underlying whole content will be lost. Hence removing the watermark is only possible by destroying the whole content. In the system the auditor will generate the data lineage which consists of the list of consumers who visited the document. It provides the means to identify the data leaker or the guilty party if any leakage occured. The main goal of the system was to identify the guilty entity. But it cannot stop the data leakage. And also requires a secure key for the encryption and it was time consuming.

In [2] Lin Yuan suggested a different and more secured solution for the privacy preserving photo sharing architecture based on JPEG scrambling technique. It can ensure privacy of the user and preserves the usability of online photo sharing services. A multi-region selective JPEG scrambling algorithm is used that can ensure photo privacy for multiple people involved in the photo. But once the scrambling is done, it requires a secure key to descramble the image.

In [3] Hasan introduced an architecture for capturing provenance of the data. Due to the increased amounts of valuable informations being produced and persist digitally, the ability to determine the origin of data becomes important. Hence data provenance tracking has become an important task for the protection of rights and authentication of information. Provenance information gives the summary of the history of the ownership of items and the different actions performed on them. This paper presents a system that implements the function of read and write actions in a tamper-proof chain. It provides the visibility of provenance information that gives the list of owners who visited the document. These provenance information is stored in the hardware. Hence the attacker can be able to strip of the information of the file so that provenance chain gets corrupted and trustness of the owner will be lost.

Multimedia security research focuses on the security of the content published, protecting the intellectual property of the content owners and creators or writers from malicious users. In [4] N. P. Sheppard presented the problem of protecting the intellectual property rights of the author from insiders, who are involved in the creative process before publication. Data generator or the creator consists of multiple single entities. One of these publishes a version of the document. All the entities involved in the generation process have access to the original document. They can publish it without giving credit to the other authors. To identify each owner separately methods of proof-of-ownership or fingerprinting can be applied. But the leakage of the document without being tracked.

In [5] Roland Parviainen suggested a scheme for scalable fingerprinting of multicast media. Each receiver of a multicast session has to be supplied a stream with unique watermark. The embedded watermarked streams helps to trace those users who make unauthorized copies of a stream. The watermarking is done by the encryption of two slightly different copies of the original stream with a set of different keys. The sender first has to split the file into blocks and then for each block he has to create two different versions. Then watermark each of them with different watermarks and encrypts each of them with different keys. Each recipient is assigned a group of keys. The resulting combination of parts can identify the recipient uniquely. The system thus can decrypt exactly one version of each part. But the problem of an untrusted sender is not addressed by this system.

In [6] Josep Domingo-Ferrer suggested the scheme for anonymous fingerprinting based on committed oblivious transfer. Anonymous fingerprinting is defined as the technique for copyright protection which is compatible with buyer anonymity in electronic transactions. This system presents the first fingerprinting protocol. It makes use of oblivious transfer that is, documents are split into smaller parts and for each part two different versions are created. The recipient can be identified by the unique combinations of versions he received. But in this system a malicious sender can offer the same version twice in the oblivious transfer. Also, it can identify which version the recipient receives. Hence the problem of data leak cannot be solved.

## IV. PROPOSED METHODOLOGY

LIME is a data lineage framework for the data flow between different entities. The entities involved in data transfer includes the owner and the consumer. In this research introduced an additional entity named auditor, whose task is to determine a guilty party for any data leakage. Auditor will be invoked only whenever a leakage occurs. The auditor then informs the owner of the document about the data leakage and will reconstruct the data lineage.

### *A. Preprocess*

Data lineage can be defined as the data life cycle. It gives the information about the transfer of data from the owner to the consumers. It describes what happens to the data as it goes through malicious environments. Data lineage construction is the creation of the list of consumers who receives the document from the owner of the document. The lineage gives the data about who are all visited that particular document. The main intention of pre-processing is to change the given document into a consistent format.

The user has to create an account for data transfer. Then the user has to login to their account for data transfer. Then the user select the document from his system or from the web. The document selected must be in JPEG format, otherwise convert it to that format. JPEG documents has high controlled degree of compression, small file size and format is compatible. It can be displayed correctly in any browsers.

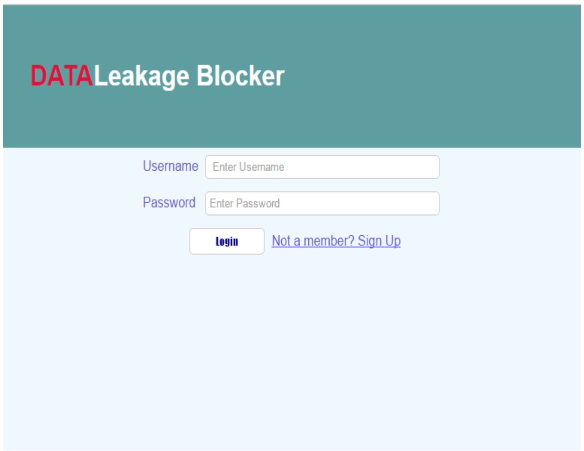
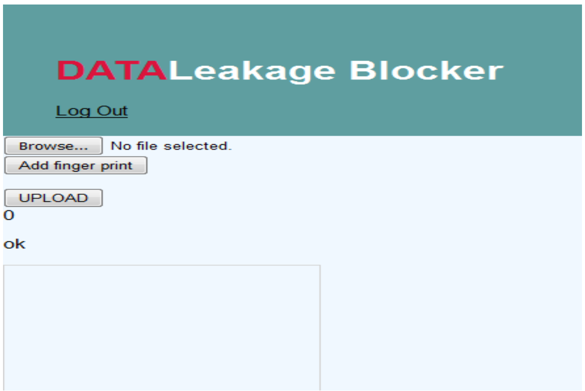


Fig. 2 Login page

The fig.2 above shows the login page. A new user can sign up an account for data transfer and then can login into the system by providing their username and password on the required area. The below fig.3 shows the home page. Data lineage will be shown on the home page itself. So that whenever the user login to their account, they can view the message that shows the details about the guilty party.



DATALeakage Blocker

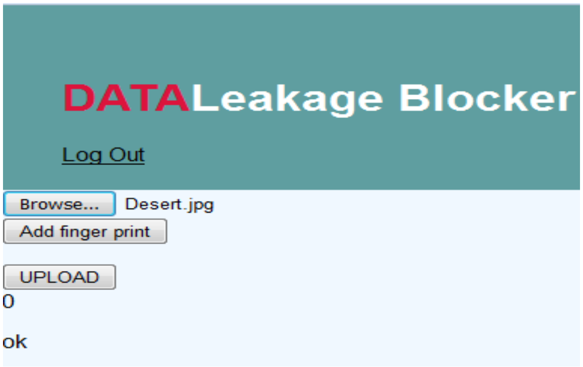
No file selected.

0

ok

Fig. 3 Home Page

The fig.3 above shows the home page. The “Browse” function is to choose a document for the data transfer. It can be selected from our system or from the web. The selected document must be in jpeg format. The “Add finger print” function is to add the embedded data so that it can uniquely identify the recipient. The “Upload” function is to upload the selected document. For example, consider that an user 1 wants to upload his document to the web. So that the other users can use the document but cannot be able to upload it again to the web. At first user 1 will login to the system using his username and password. Then he select the document to upload from his system.



DATALeakage Blocker

Desert.jpg

0

ok

Fig. 4 Desert.jpg

The fig.4 above shows the document selected named desert using the browse function. The Jpeg documents can only be selected as it is compatible with any browser. Then the user has to provide security to the document. The fingerprinting is applied for this purpose.

### B. Enhanced Least Significant Bit Algorithm (ELSB)

To provide the fingerprint to a document the Enhanced least Significant Bit algorithm is used [7]. The algorithm defines as follows,

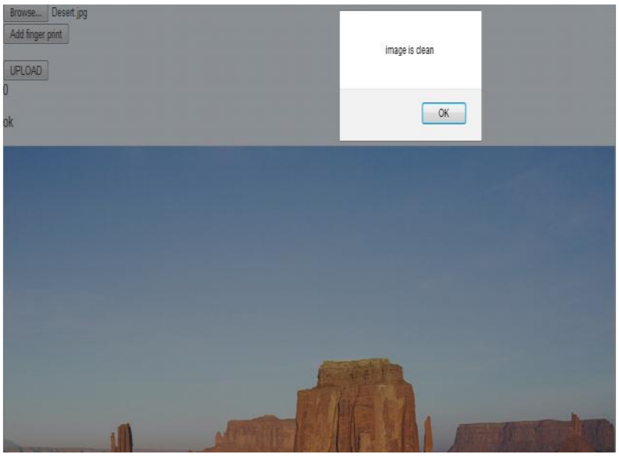
- Select a cover image of size M*N as an input.
- The message to be hidden is embedded in any color of RGB component of the image.
- Then use a pixel selection filter to obtain the best areas to hide information.
- The filter is applied to Enhanced Least Significant Bit of every pixel to hide information.
- The message is hidden using Bit Replacement method.

ELSB [7] minimizes distortion level which is negligent to human eye. It improves the performance of LSB by hiding data in any of the three colors.

### C. Add Fingerprint

The most important step is to add fingerprint to the selected document. The owner of the document has to embed some information that uniquely identifies the recipient. This method of adding fingerprint is termed as fingerprinting. Whenever a document is transferred to a consumer, the owner must add fingerprint. If the consumer leaks this document, it is possible to identify him with the help of the embedded information namely fingerprint.

For example if we click the “Add finger print” button in fig.4 a message box will be displayed as in fig.5 which defines the status of the document. The status includes whether the document is clean or already fingerprinted.



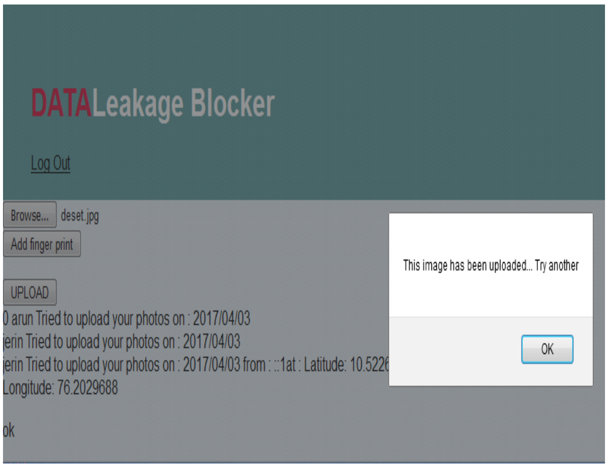
Desert.jpg

0

ok

image is clean

Fig. 5 Image is clean



DATALeakage Blocker

deset.jpg

0 arun Tried to upload your photos on : 2017/04/03
jerin Tried to upload your photos on : 2017/04/03
jerin Tried to upload your photos on : 2017/04/03 from : ::1at : Latitude: 10.5226
Longitude: 76.2029688

ok

This image has been uploaded... Try another

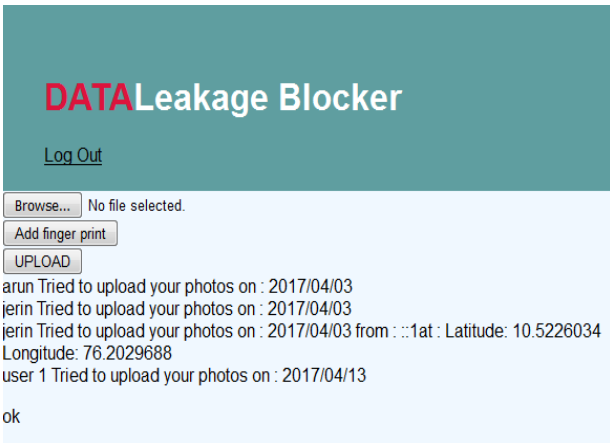
Fig. 6 Image was already uploaded

The fig.5 above shows the message that the image is clean. It defines that the document is pure and it has no embedded data. Hence the user can upload the document by using the upload function. Then the user has to save the uploaded fingerprinted document. The message about the document fingerprint will help the honest consumers to protect themselves from data leak. If a particular document was already fingerprinted and uploaded, then a message will be displayed which shows that the "image was already uploaded". Hence honest parties can save themselves by not uploading that particular document.

In the second scenario, if any user other than the document owner tried to add fingerprint to the document which was already fingerprinted by the document owner itself, then also a message box will be displayed. The message box will displays a message that "This image has been uploaded..Try another", as shown in fig.6 above. Hence the new user cannot make changes or to add fingerprint to that document.

### D. Auditing

A key position of LIME is given to the auditor. Auditor monitor the communication between the owner and the consumer. He is not involved in the data transfer, but he takes action once a leakage occurs. If the owner found that his document has leaked he will invoke the auditor and provide the leaked data. The auditor will analyze the document and construct the data lineage. Data lineage generated consists of the list of consumers who tried to upload the owner's document. Then the auditor will give the details about the unauthorized user to the document owner. Hence the owner can identify the leaker of their document.



```
arun Tried to upload your photos on : 2017/04/03
jerin Tried to upload your photos on : 2017/04/03
jerin Tried to upload your photos on : 2017/04/03 from : ::1at : Latitude: 10.5226034
Longitude: 76.2029688
user 1 Tried to upload your photos on : 2017/04/13

ok
```

Fig. 7 Data Lineage

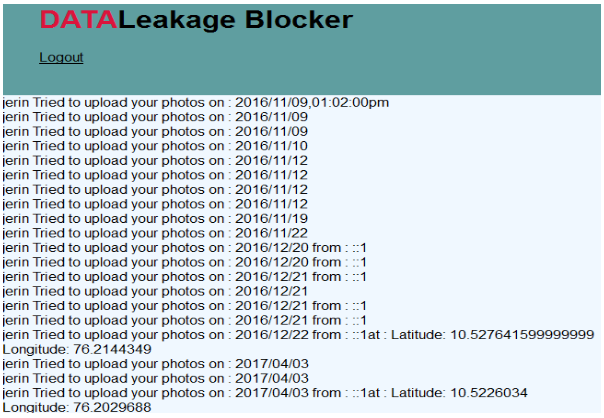
For example, if user 1 tried to upload the document of user 2 by clicking on the upload button in fig.4, he will be directly sent to the login page. Thus the data leakage can be blocked. Then that user cannot do any task using that document. Then the user 1 has to login again and continue his task. Thus the system will blocks the data leakage and then informs the owner about the data leak.

The auditor will be invoked by the owner of the document to save the identity of the user. By using the embedded information called fingerprint the auditor can create an ordered chain of consumers who received the document. Hence user 2 will get the message from the auditor which contains the details about the user 1 as shown in fig.7 below. Hence the owner can easily detect the leaker of the document from the message given by the auditor which will be displayed at the home page itself of the owner.

The fig.7 above shows the lineage of the leaked document. Whenever any of the user other than the document owner tried to upload the document, auditor will be invoked. Hence auditor will block the data leak and then informs the owner about the leaker. Thus the owner will receives the message about the data leak at his home page itself.

### E. Auditor Login

The auditor of the system can be any authority, for example it can be a governmental institution, police, a legal person or even some software. To provide more trustness to the auditor, the auditor login has done which makes the system more secure. So that only authorized auditor can be involved in data lineage construction.



```
jerin Tried to upload your photos on : 2016/11/09,01:02:00pm
jerin Tried to upload your photos on : 2016/11/09
jerin Tried to upload your photos on : 2016/11/09
jerin Tried to upload your photos on : 2016/11/10
jerin Tried to upload your photos on : 2016/11/12
jerin Tried to upload your photos on : 2016/11/12
jerin Tried to upload your photos on : 2016/11/12
jerin Tried to upload your photos on : 2016/11/12
jerin Tried to upload your photos on : 2016/11/19
jerin Tried to upload your photos on : 2016/11/22
jerin Tried to upload your photos on : 2016/12/20 from : ::1
jerin Tried to upload your photos on : 2016/12/20 from : ::1
jerin Tried to upload your photos on : 2016/12/21 from : ::1
jerin Tried to upload your photos on : 2016/12/21
jerin Tried to upload your photos on : 2016/12/21 from : ::1
jerin Tried to upload your photos on : 2016/12/21 from : ::1
jerin Tried to upload your photos on : 2016/12/22 from : ::1at : Latitude: 10.527641599999999
Longitude: 76.2144349
jerin Tried to upload your photos on : 2017/04/03
jerin Tried to upload your photos on : 2017/04/03
jerin Tried to upload your photos on : 2017/04/03 from : ::1at : Latitude: 10.5226034
Longitude: 76.2029688
```

Fig. 8 Auditor Login

In the above fig.8 shows the auditor login. The auditor constructs the list of the unauthorised users who tried to upload the documents. Auditor displays the data lineage and also inform the owners about data leakage. The auditor detects the data leaker by identifying his name and location and then sent those details to the owner of the document. Hence the owner can detect the data leakage and identify the data leaker from the data provided by the auditor.

## V. EXPERIMENTAL RESULT

The hardware requirements of the system includes :

- Processor - Pentium IV 2.4 GHz
- RAM - 2 GB
- Hard Disk – 350 GB

The software requirements of the system includes :

- Operating System - Windows 7 and above
- Programming language - PHP and Javascript
- Server – Apache
- Database - MySQL

HTML5 Base64 canvas is used for the binary to text encoding. The HTML5 canvas element provides to draw graphics on web pages. It acts as a container for the documents. JavaScript is used to draw the graphics. Only JPEG documents are compatible with the system. Hence document selected must be in JPEG format, otherwise convert it to that format. JPEG documents has high controlled degree of compression, small file size, and user can independently select the ratio of quality or file size [2]. It can be displayed correctly in any browsers and also any of the devices like computers, tablets and mobile devices.

In the existing system the fingerprinting was done by using watermarks [1]. Watermarked documents are easy to identify and a malicious user can be able to remove the watermarks and can recover the embedded content in it. Hence inorder to make it more secure watermarking was done on the core part of the document. Hence if the watermark is removed the whole content will be lost [8]. The unauthorized user can have a tendency to detect, intercept and modify the watermarked images. Hence to provide more security to the documents steganography is used for fingerprinting in this system. The steganographic documents are difficult to detect, as there won't be any visible changes to the actual document. Steganographic documents does not allow any enemy to even detect the presence of a second message hidden in that document. It can transfer the data securely to the destination without any modifications.

The watermarking technique seems to be more time consuming [9] than steganography and also the image clarity will be lost. But steganograhic fingerprinting is less time consuming and clarity of the document can be maintained and also it is not possible to detect the fingerprinting with the naked eye. In watermarking we have to consider the neighbouring pixels values [10] also making it more complex whereas in steganography it is not considered. The maximum size of a watermark can have is limited by the size of the document. This may be difficult for documents of small sizes.

The ELSB algorithm [7] provides minimum distortion level which will be negligent to human eye. It improves the performance by hiding data in any of the three colors. This makes it difficult to detect with human eye. The ELSB algorithm can be evaluated by using imperceptibility parameter to calculate the similarity of the original document and the stego document.

### A. *Imperceptibility*

The imperceptibility is the measure of similarity between the original document and the stego document. When the human eye cannot distinguish between the original image and the stego image then the process can be said to be imperceptible. This parameter can be evaluated by using the PSNR values. The greater the value of PSNR shows the greater degree of imperceptibility. Hence the distortions in the stego document cannot be detected by human eye. The ELSB algorithm shows a higher PSNR value when compared with the LSB algorithm. Hence the ELSB algorithm is highly imperceptible.

To ensure the image quality MSE and PSNR values are calculated.

### B. *Mean Square error (MSE)*

MSE is the Mean Square error which gives the rate of distortion between the original and stego image in decibels. The more the MSE values the more will be distortion which means that the difference between the cover image and stego image is more. The lesser the MSE lesser will be the distortion which means that difference between the original image and the stego image is very less.

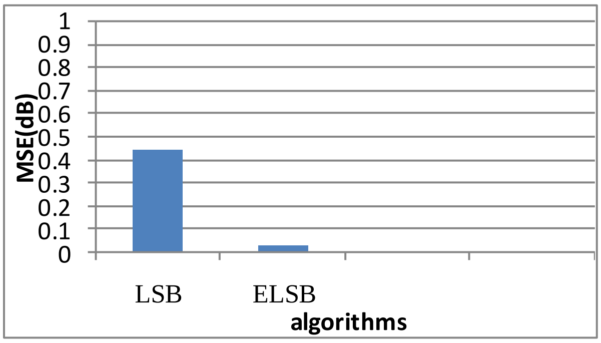


Fig. 9 MSE comparison

The above fig.9 shows the comparison of MSE values between LSB and ELSB methods. The value of LSB is 0.4405 dB and ELSB is 0.027 dB. The more the MSE values the more will be distortion. In the above figure LSB shows more MSE value and it has more distortion compared to ELSB. The low value of ELSB shows that it has only less difference between original and stego images. Hence ELSB has the minimum distortion.

### C. *Peak Signal to Noise Ratio (PSNR)*

PSNR is the Peak Signal to Noise Ratio which represents the proportion of distortion between the cover image and the stego image in decibels. The more the PSNR values the less will be distortion. The larger PSNR values indicates that there is more similarity with the cover and stego images and less distortion.

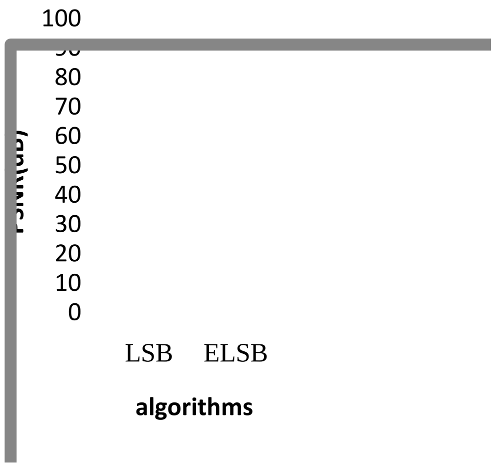


Fig. 10 PSNR comparison

The above fig.10 shows the PSNR values of the LSB and ELSB technique. The larger the PSNR values the larger is the clarity. In the figure ELSB shows more PSNR value. Large PSNR values states that only a small difference between cover and stego images. It shows that ELSB provides the minimum distortion.

Table.1. Comparison of data lineage detection

| **Existing System** | **Proposed System** |
|---|---|
| Watermarking | Steganography |
| Easy to identify watermark | Difficult to detect steganograhic document |
| Changes to the original image was visible | No change to the original image |
| Watermarking is time consuming as it has to consider neighbouring pixel values | Less time for steganography. Neighbouring pixel values are not considered |
| Document can be uploaded by consumers | Document can't be able to upload by the consumers |
| Data lineage generation after the data leakage | Data lineage generation before the data leakage itself |
| Cannot block data leakage | Can block data leakage |
| Auditor was not honest | Auditor is honest |
| Less secure | More secure |

The Table.1 above shows the comparison of the existing system with the proposed system. From the above comparison it is clear that the proposed system is the more better solution for data lineage detection. It can construct the data lineage before the data leak and informs the owner about the leaker whenever a leak occurs. The system is more secure and efficient as it can block the data leakage.

## VI. CONCLUSION AND FUTURE SCOPE

LIME is a model for data transfer between multiple entities. This system provides a framework for accountability by design. It enables the honest parties to provide the needed protection for their sensitive data. In the existing system data lineage generation was done only after the data leakage. But in our system data lineage will be generated before the data leak itself. This system will detect the guilty entity and blocks the data leakage. The system will provide more security to our documents by blocking its leakage and also detects the guilty party who tried to leak the document. Hence our system is more efficient and secure. The future scope of this research includes the data lineage detection for all other types of documents.

## REFERENCES

[1] Michael Backes, Niklas Grimm, and Aniket Kate, "Data Lineage in Malicious Environments,"in *IEEE Transactions on Dependable and Secure Computing*, vol.13, issue.2, 2016.

[2] Lin Yuan, Pavel Korshunov, and Touradj Ebrahimi, " Secure JPEG Scrambling enabling Privacy in Photo Sharing, " *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG),* Vol: 04, 2015.

[3] R. Hasan, R. Sion, and M. Winslett, "The case of the fake picasso: Preventing history forgery with secure provenance," in *International Confr* , 2009.

[4] N. P. Sheppard, R. Safavi-Naini, and P. Ogunbona, "Secure multimedia authoring with dishonest collaborators," *EURASIP J. Appl. Signal Process*., vol. 2004

[5] R. Parviainen and P. Parnes, "Large scale distributed watermarking of multicast media through encryption," in *Proceedings of the IFIP TC6/TC11 International Confr.* vol. 192, 2001.

[6] J. Domingo-Ferrer, "Anonymous fingerprinting based on committed oblivious transfer," *in Public Key Cryptography*. Springer, 1999.

[7] Shilpa Guptha, Geeta Gujral, and Neha Aggarwal, "Enhanced Least Significant bit algorithm for image steganography," *in IJCEM,* Vol. 15, Issue 4, July 2012.

[8] J. P. M. Linnartz and M. Van Dijk, "Analysis of the sensitivity attack against electronic watermarks in images," *in Information Hiding*. Springer, 1998.

[9] J. T. Brassil, S. Low, and N. F. Maxemchuk, "Copyright protection for the electronic distribution of text documents," *Proceedings of the IEEE*, vol. 87, no. 7, 1999.

[10] I. J. Cox and J. P. M. Linnartz, "Public watermarks and resistance to tampering," *in International Conference on Image Processing,* 1997.