Data Leakage Detection using Image and Audio Files

P.P. Dandavate
Department of Computer Engineering,
Bharati Vidyapith Deemed University College of Engineering,
Pune

S.S. Dhotre
Department of Computer Engineering,
Bharati Vidyapith Deemed University College of Engineering,
Pune

ABSTRACT
Accidental or intentional distribution of data to unauthorized entity is the data leakage .In business process, it is necessary to send sensitive data to trusted parties. But this data is found at unauthorized place such as website or somebody's laptop. It is very challenging and important to detect leakage when sensitive data is deliberately leaked to others. Traditionally leakage detection is handled by watermarking technique. But it involves modification of data. In this paper for accessing "guilt" of agent a model is developed. Algorithms are presented to distribute objects in such a way that increase the chances of detecting leaker. Finally fake object is included in distributed set using steganography LSB algorithm which do not modify individual members. For the entire set fake object is acting as type of watermark . Major contribution to this system is to develop guilt model using steganography LSB algorithm.

Keywords
Data leakage, fake objects, Guilt Model

1. INTRODUCTION
Security to data is important when data is given to trusted third parties. In business process, sensitive data is shared among various employees, business partners and customers. Sensitive data include financial information; patient information and other information depending on the business and industry.Somtimes Company have partnership with other company that requires sharing customer data. There is possibility of leakage of data. In this system owner of data is called distributor and trusted parties as agents. Goal of this system is to find which data of distributor's has been leaked and if leaked detect agent who leaked data.

Traditionally for leakage detection watermarking technique was used .In that unique code is included in distributors' data

If that data is found at unauthorized place leaker can be identified. But watermarking technique involves modification of data. Also if data receiver is malicious watermarks can be destroyed.

In this system applications are considered where original sensitive data cannot be perturbed. In Perturbation technique data is modified and made less sensitive before being handed to agents. For example, one can add random noise to certain attributes or one can replace exact values by ranges .In this paper following scenario is discussed: Distributor find the set of object at an unauthorized place after giving set of object. At this point distributor can assume that data has leaked by agents instead of gathered by other means.

In this system a model is developed for finding "guilt" of agent. To increase the chances of detecting agents algorithms are used. Finally before being giving data to agents fake objects has been added to distributor set using steganography LSB algorithm.

2. RELATED WORK
Approach that is used for detection of guilty agents is similar to data provenance problem. Data allocation strategies used in this system which was used initially in images videos and audio data is similar to watermarking technique was used to maintain original ownership of distributed data. Approach that is used in this system and watermark is same i.e. providing agents with some kind of receiver identifying information. But watermarking technique modifies the item being watermarked. If object that is to be watermarked cannot be modified then watermark cannot be inserted into object. In such case methods that use watermark to the distributed set is not applicable.

3. PROPOSED WORK
In this system, model is developed for accessing "guilt" of agents. Algorithms are used for distributing objects in such a way that increase the chances of detecting agents. Finally while distributing objects to agents fake objects are added using steganography LSB algorithms. This is the main contribution of this system. The system architecture is shown in fig 1.

The system gives access to data distributor as well as agents registered by data distributor.

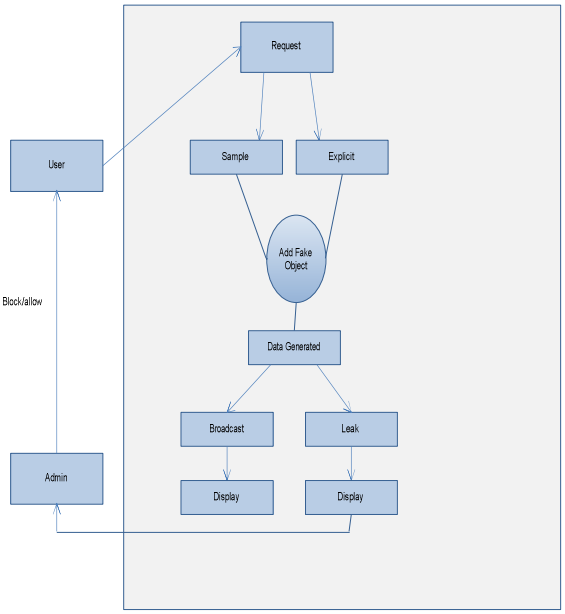


Fig.1 System Architecture

Access to the system is given only to the registered agents. Requested data can be accessed by the agents. If data is leaked and distributor finds that data at unauthorized place distributor can send this data to system and system finds guilty agents.

Total workflow of system is divided into four modules.

Module 1
In this module design for the website as well as the authorized agent will logged in. For new agent has to fill the register form, after successfully submitting the of data he can logged into our system.

Module 2
In this module the agent request for the data, he will get the data from server as per request by adding fake object. Using steganography LSB algorithm.

Module 3
In this module user has two options broadcast and leak the data. If user selects leak option then the unauthorized news channel will be updated. And if user selects the broadcast option then authorized news channel will be updated.

Module 4:
This module checks whether our system data as well as another system data is same or not. Or in short find out who is guilty agent and take action on guilty agent (put him in black list). Administrator module for approve and decline the new register user.

3.1 Steganography LSB algorithm:

This is the simplest steganography method based on the use of LSB. It consist of sequential substitution of each Least significant bit of image pixel for bit message. The following method illustrate how to hide the message "A" in cover image

1st step :Data is converted form decimal to binary.

2nd step: Cover image is read as shown in fig 2.

Fig.2 The cover image

144	142	146	152	156	147	151	157
160	155	159	162	133	123	133	145
144	141	141	138	61	55	65	79
120	123	131	144	50	61	74	92
170	167	167	166	61	59	56	59
120	125	131	132	61	59	59	59
124	133	139	131	88	76	77	76
138	153	167	154	139	.....	.....	.....

3rd step: Image is converted from decimal to bianry.

10010000	10011010	10011100	10010010	10010110	10011101	10101111	10100101
10100000	10011011	10011111	10100010	10000101	01111011	10000101	10010001
10010000	10001101	10001101	10001010	00111101	00110111	01000001	01001111
01111000	01111011	10000011	10010000	00110010	00111101	01001010	01011100
10101010	10100111	10100111	10100110	00111101	00111011	00111000	00111011
01111000	01111101	10000011	10000100	00111101	00111011	00111011	00111011
01111100	10000101	10000111	10000011	01011000	01001100	01001101	01001100
10001010	10011001	10100111	10011010	10001011	.....	.....	.....

4th step:byte is divided into hidden bits.

Thus [1|0000001] is divided into 8 bits [1 0 0 0 0 0 0 1].

5th step:First byte of original data is taken from the cover image.

10010000	10011010	10011100	10010010	10010110	10011101	10101111	10100101

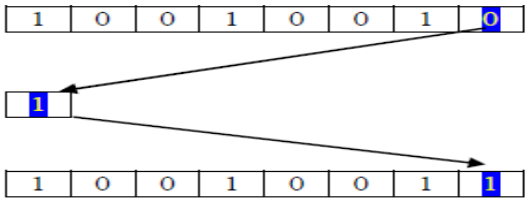
First byte of original data from cover image

1	0	0	1	0	0	0	0

-first bit of data which is to hide:

1

Change the least significant bit



-Repeat same procedure for all bytes of cover image.

Before and after steganography cover image is shown in fig.

4. PROBLEM SETUP AND NOTATION

4.1 Entities and Agents

A distributor has set T= {T1, T2…TM} objects. The distributor distributes objects to agents U1, U2,….Un .Objects could be image or MP3 file.

Fig.3: Cover Image before steganography

Fig. 4: Cover Image after steganography

Any agent receives subset of objects either by sample request or explicit request.

Sample Request Ri ==Sample (T, mi) Any subset of mi objects is given to from T can be given to Ui**.**

Explicit request: Explicit (T, condi) In this request agent Ui receives objects that satisfy condition.

4.2 Guilty Agents

After giving subset of objects distributor finds that some of the S objects were leaked. Since distributor has given data to agents U1, U2….Un. Distributor can suspect that agent is leaking the data. But agents can specify they have not leaked data and from other means data is leaked.. Goal of this system is to find that leaked data came from agents instead of other means.

While distributing data to agents, unique fake object is included in agents data .If system finds leaker, it can been said that that agent is guilty agents.

5. DATA ALLOCATION ALGORITHM

Distributor "intelligently distributes data to agents in order to improve chances of detecting guilty agents. Two types of request were made by agents, sample & explicit. Distributor *generates fake objects which are not in set T. There is no difference between fake object and real object*

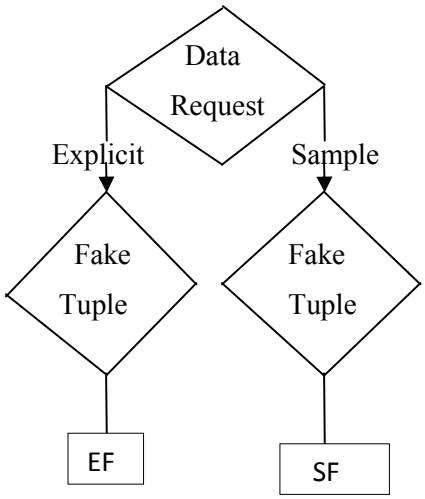


Fig 5: Leakage instances

5.1 Explicit Data Request

Explicit Data request

In this allocation strategy, agent request data on a constraint i.e. distributor has to distribute data satisfying some condition.

Algorithm 1 Algorithm for Explicit Data request (EF)

Input: P1... Pn, cond1... condn,c1..c2, S

Output: P1...Pn, F1... Fn

1: R ; . Agents that can receive fake objects

2: for i ¼ 1... n do

3: if ci > 0 then

4: P← P U{i}

5: Fi←∅ ;

6: while S > 0 do

7: i←SELECTAGENT(P, P1 …Pn)

8:f←CREATE FAKEOBJECT (Pi, Fi,condi)

9:Pi←Pi U {f}

10:Fi←Fl U {f}

11.ci←ci-1

12:if ci=0 then

13 :P←P\{Pi}

14:S←S-1

5.2 Sample Data Request

Algorithm 2 Allocation for sample Data Request (SF)

In this request distributor distributes any subset of data to agents. Sample request Ri = From T any subset of mi records can be given to Ui.

1: a ← 0|T| a[k]: number of agents who have received object tk

2: R1 ← □, . . . ,Rn←□

3: remaining ← Σ*i*= 1 *n mi* // No of sample sets that we have to distribute to agents

4: while remaining > 0 do

5: for all i = 1, . . . , n : |Ri| < mi do

6: k ← SELECTOBJECT(i,Ri) May also use additional Parameters

7: Ri ← Riυ {tk}

8: a[k] ← a[k] + 1

6. EXPRIMENTAL RESULTS

System flow is as follows.

1. Agent request either EXPLICT or IMPLICT request

2. Accoring to request username is embedded in image or audio file.

3. If leakage is found username is extracted from system and it is compared with all users in the system.

4. If it is same as that of user leaker can be found.

Figure 6 below shows space comparison between image and audio file. After embedding username how much space that usename take that comparison is shown

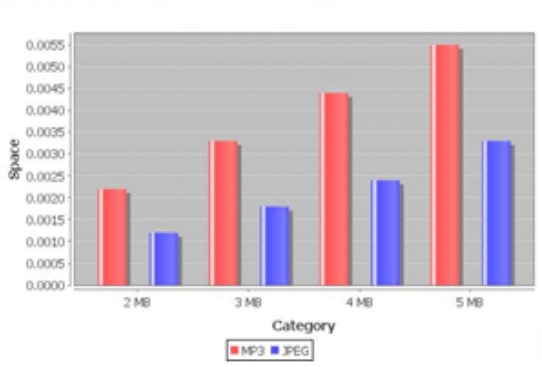


Fig 6: Space comparison graph

Figure 7 shows time comparison between image and audio file. For embedding username how much time image file and audio file requires That is shown by graph.

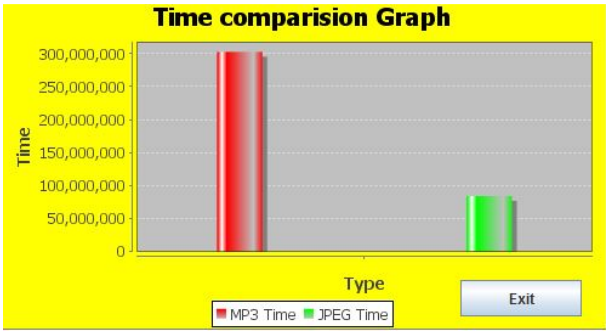


Fig 7: Time comparison graph

7. CONCLUSION

Everyday data leakage happens when confidential business information is leaked out. It is not certain that leaked data came from agent or some other means. If we have to handover sensitive data we could watermark it so that we could trace out its origin. But some data cannot include watermark.

Considering all these difficulties, this system is showing that agent is leaking data . The algorithms that are used distribute the data in such a way that it increases the chances of detecting leaker. While distributing data to agents fake object are added using steganography LSB algorithm. It is the main contribution of this system

8. REFERENCES

[1] A. E.Mustafa A.M.F.ElGamal M.E.ElAlmi Ahmed. BD" A Proposed Algorithm For SteganographyIn Digital Image Based on Least Significant Bit"

[2] F. Hartung and B. Girod, "Watermarking of Uncompressed and Compressed Video," Signal Processing, vol. 66, no. 3, pp. 283-301, 1998

[3] J.J.K.O. Ruanaidh, W.J. Dowling, and F.M. Boland, "Watermarking Digital Images for Copyright Protection," IEE Proc. Vision, Signal and Image Processing, vol. 143, no. 4, pp. 250-256, 1996.

[4] L. Sweeney, "Achieving K-Anonymity Privacy Protection Using Generalization and Suppression,"http://en.Scientificcommons.org/43196131, 2002.

[5] P. Buneman, S. Khanna, and W.C. Tan, "Why and Where: A Characterization of Data Provenance," Proc. Eighth Int'l Conf. Database Theory (ICDT '01), J.V. den Bussche and V. Vianu, eds.,pp. 316-330, Jan. 2001.

[6] Papadimitriou and H. Garcia-Molina "Data leakage detection "IEEE Transaction on knowledge and data engineering, pages 51-63 volume 23, January 2011.

[7] S. Czerwinski, R. Fromm, and T. Hodes, "Digital Music Distribution and Audio Watermarking, "http://www.scientificcommons.org/43025658, 2007