Comparative Analysis of Database Forensic Algorithms

Manish H. Bhagwani
Asst. Prof. (CSE)
Chhattisgarh Institute of Technology,
Rajnandgaon,CG, India

Rajiv V. Dharaskar, PhD
Director
Disha Education Society,
Raipur, CG, India

V. M. Thakare,PhD
Professor & HOD
PG Dept. of CSE,
SGBAU, Amravati, MS, India

ABSTRACT
Security is major concern in data outsourcing environment, since data is under the custody of third party service provider in many cases. In present systems, third partycan access & view data even though they are not authorized todo so or even when the data is outsourced to the auditors orallow the employee of the organization to do the updating in the database. This may lead to the serious data theft, data tampering & even data leakages causing severe business impact to data owner. Digital Forensic analysis of databases helps to solve the problem. In this paper various database tamper detection algorithms are studied and compared based on space and time complexities.

Keywords
Security, Data outsourcing, Forensic Analysis, Digital Forensic, Tamper Detection Algorithms.

1. INTRODUCTION
Secure data storage is an everyday requirement for public businesses, government agencies and many institutions. For many organizations, if data were to be maliciously changed, whether by an outsider or by an inside intruder, it could cause severe consequences for the company. Possibly even for their clients as well. There are many reasons why someone might want to tamper with data. For example, an unsatisfied student who receives a "D" grade in his mathematics subject, in which he needed at least a "B", could be highly tempted to try to dishonestly change his grade to a "B" in the school'sdatabase. This would be an example of someone who would have to hack into the system from outside, unless of course the student somehow had access to the database containing the grade.

A similar example, wherein the intruder is an insider rather than someone hacking in from the outside, could be that ofan employee at a large company who is trying to meet his sales requirements for a fiscal year. He might attempt to change the dates of transactions to make it appear that they had transpired within the previous fiscal year when, inreality, they had not.

Data outsourcing is an emerging paradigm that allows users & companies to give their (potentially sensitive) data to the external servers that then become responsible for the storage, management and dissemination.

By outsourcing organization can concentrate on their core business activity rather than incurring the substantial hardware, software and personnel cost involved in maintaining applications in-house. Although data outsourcing provides many benefits especially for parties with limited data operators for managing an ever more increasing amount of data, it also introduces new privacy and security concerns.

As promising as it is, this paradigm also brings many new challenges for data security. When business organizations outsource sensitive data for sharing on servers, which are not within the same trusted domain as data owners, in data outsourcing scenario, access to data is selective, with different users enjoying the different views over the data. When the data is outsourced there is therefore the problem of enforcing possible data theft or data tampering by theinside employees or by the third party data auditors or it may be from any other form of internal or external threats. In the data outsourcing scenario the data operators areunder the strict custody of trusted party which monitorseach access request to verify if it is compliant with the specified client or not. This approach requires some additional measures to beconsidered. There is need for data owner (business organization) to manage access to legitimate users. To achieve this, owners can use digital signatures to identify the persons for whom they allow to access data. This actually leads to a system called notarization, which is been used by another system called validator to check for the data redundancy and data originality.

Outsourcing healthcare Insurance services is extremely popular today. However, there are several concerns being voiced about data security and adhering to standard quality norms. The Health Insurance Portability and Accountability Act (HIPAA) are widely acknowledged as the norm for healthcare services and Indian companies are well versedwith the Act and other regulatory bodies. Some other standards/acts relevant for data security are:

- The Information Technology Act 2000 (ITA-2000),
- Payment Card Industry Data Security Standard (PCIDSS)
- ISO 27001, ISO27001 Information Security Standard

2. RELATED WORK
Widespread news coverage of collusion between auditors and companies they audit [1], a recent FBI study indicates that almost half of attacks were by insiders [2].It is assumed that the notarization and validation services remain in a trusted computing base. This can be done by making them geographically and perhaps organizationally separate from the DBMS and the database [3], thereby effecting correct tamper detection even when the tampering is done by highly motivated insiders. Scenario, like discusses tampering event in which in U.S., all patients are required to sign an authorization under HIPAA [4]. Computer forensics is now an active field, with more than 50 books published in the last 10 years. There are few computer tools for these tasks, in part due to the heterogeneity of the data. One substantive example of how computer tools can be used for forensic analysis is Mena's book [5]. Goodrich et al. introduce new techniques forusing main-memory indexing structures for data forensics [6].

In the database context, previous papers introduced the approach of using cryptographic hash functions to detect database tampering [7] and of introducing additional hash chains to improve forensic analysis [7]. Previously, there has been proposed the Monochromatic, RGB, and Polychromatic forensic analysis algorithms [8].

If an adversary modifies even single byte of data or itstimestamp, the independent validator will detect a mismatch with the notarized document, thereby detecting the tampering. The adversary could simply re-execute transactions, making

whatever changes he/she wanted, and then replace original database with his/her altered one. However, the notarized document would not match in time. Avoiding tamper detection comes down to inverting the cryptographically strong one way hash function. An extensive presentation of an approach, performance limitations, tamper detection, threat model and other forensic analysis algorithms is discussed in paper [7],[9]. Hash chain linking is discussed in more detail in paper [7].

Tiled bitmap algorithm is refinement of polychromatic algorithm. The advantage of the Tiled Bitmap Algorithm isthat it lays down a regular pattern (a "tile") of such chains over contiguous segments of the database. The other advantage of the Tiled Bitmap Algorithm is that it can detect multiple corruption events that other previous algorithms cannot. On the other hand it suffers from false positives while the previous algorithms do not.

2.1 Monochromatic Algorithm

The Monochromatic Algorithm uses only the cumulative (black) hash chains and as such it isthe simplest algorithm in terms of implementation.

2.2 RGB Algorithm

In the RGB Algorithm, three new types of chains are added, denoted with the colors red, green, and blue, to the original (black) chain in the so-called Monochromatic Algorithm. These hash chains can be computed in parallel; all consist of linked sequences of hash values of individual transactions in commit order. While additional hash values must be computed, no additional disk reads are required. The additional processing is entirely in main memory. The RGBY Algorithm retains the red, green, and blue chainsand adds a yellow chain. This renders the new algorithm more regular and more powerful.

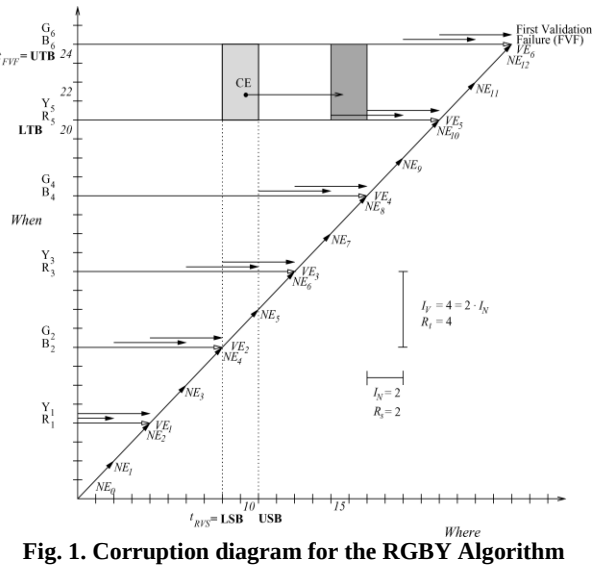

Fig. 1. Corruption diagram for the RGBY Algorithm

2.3 RGBY Algorithm

The RGBY Algorithm is an improvement of the original RGB Algorithm. The main insight of the previously presented Red-Green-Blue forensic analysis algorithm (or simply, the RGB Algorithm) is that during notarization events, in addition to reconstructing the entire hash chain (Illustrated in Fig 1. with the long right-pointed arrows in prior corruption diagrams), the Validator can also rehash portions of the database and notarize those values, separately from the full chain.

2.4 A3D Algorithm

The a3D Algorithm is the most advanced algorithm in the sense that it does not lay repeatedly a "fixed" pattern of hash chains over the database. Instead, the lengths of the partial hash chains change (decrease or increase) as the transaction time increases, in such as way so that at each point in time a complete binary tree (or forest) of hash chains exists on top of the database. This enables forensic analysis to be speed up significantly.

In all the above mentioned algorithms they differ in the amount of work necessary during normal processing. As we seen in Monochromatic algorithm we use an array Black Chains of Boolean values to store the results of validation during forensic analysis.

Computing additional hash chains during periodic validation) and the precision of the when and what estimates produced by forensic analysis.

The Boolean results are indexed by the subscript of the notarization event considered: the result of validating isstored at a given index. Since we do not wish to pre-compute all this information, the validation results are computed lazily, i.e., whenever needed. This can give rise to corruption easily.

The RGBY Algorithm was designed so that it attempts to find more than one Corruption Event. However, the main disadvantage of the algorithm is that it cannot distinguish between three contiguous corruptions and two corruptions with an intervening notarization interval between them. The a3D Algorithm is working on the recursive pattern for the call of notarization service. Where if the Chain is having lager tree then it performs faster but fails to get desired result for all the intervals.

2.5 Tiled Bitmap Algorithm

This algorithm introduces the notion of a candidate set (all possible locations of detected tampering(s)) and provides acomplete characterization of the candidate set And its cardinality. An optimal algorithm for computing the candidate set is also presented. Finally, the implementation of the Tiled Bitmap Algorithm is discussed, along with a comparison to other forensic algorithms in terms of space/time complexity and cost.

Where candidate Set Function is to arrange values of targeted binary array in reverse order and renumber function is to re arrange values of targeted binary array in perfect order.

So in our proposed System the DBMS computes a cryptographically strong one-way hash function for each tuple inserted and then notarizes it using a notarization service. This made it possible to check the consistency ofthe data by comparing it to the values stored with the notarization service. In continuation with this method, algorithms were designed to further analyze an intrusion of a database.

```
// input: Tfvf is the time of first validation failure
// IV is the validation interval
// k is used for the creation of Ct,k
// Rs is the spatial detection resolution
// output: Cset, an array of binary numbers
function Tiled Bitmap(Tfvf , IV , k, Rs)
1: t ← 0 // the target
2: Cset ← Ctemp ← .
3: T ← 1
4: while T< TFV F do
5: if ¬ val check(c0(T)) then
6: n ← lg(IV /Rs)
7: for i ← n to 1
8: t ← t + 2n.i·val check(ci(T))
9: Ctemp ← candidateSet(t, n, k)
```

10: **for each** r ∈ Ctemp
11: g ← renumber(r, T, Rs)
12: Cset ← Cset ∪ {g}
13: T← T+ IV
14: **return** Cset

Table 1: Running Time Complexity of Forensic Analysis Algorithms

Table 2: Worst Case Cost/Space Complexity of Forensic Analysis Algorithms

Algorithm	Cost (Rs=1)
Algorithm	Cost (Rs=1)
Monochromatic	O(log(D/Iv))
Monochromatic	O(D)
RGB	O(D/Iv)
RGB	O(D)
Tiled Bitmap	O((D.log(Iv)/Iv)+D)
Tiled Bitmap	O(D.(1+Ig.Iv)/Iv)

Table 3: Comparing Various Database Forensic Algorithms

Algorithm	Running Time	Cost	Space Complexity	Dynamic Performance Based on Hash Chains	Multipple Corruption Events
Monochro-matic	Fast	High	More	No	No
RGB	Low	High	More	No	No
a3D	Low	Medium	Medium	No	Yes
Tiled Bitmap	Medium	Low	Less	No	Yes

3. EVALUATION

3.1 Running Time Complexity of Forensic AnalysisAlgorithms

Table 1 shows the running time for three of the forensic analysis algorithms (the Polychromatic Algorithm isomitted because it is replaced by the Tiled Bitmap Algorithm) along with our approach. It is assumed that the spatial detection resolution Rs is equal to 1 for simplicity and D denotes the Number of Multiple corruption events. Observe that the algorithms become progressively slower because of the increased number of chains maintained and used during forensic analysis. The Monochromatic Algorithm, while being the fastest algorithm, suffers from the fact that only the first corruption event can be detected. As noted, the Tiled Bitmap Algorithm can be slightly optimized by retaining the cumulative chain of the Monochromatic in order to locate the first corrupted tile by performing binary search.

3.2 Worst Case Cost / space Complexity of ForensicAnalysis Algorithms

Table 2 shows the cost for each of the forensic algorithms assuming a spatial detection resolution of one hour (Rs=1) and a single corruption event.

In this case, we observe the opposite trend compared to theone observed for the running times of the algorithms. For a sufficiently large validation interval IV, the Tiled Bitmap Algorithm has the smaller cost. This is because the ratio (1+lgIV)/IV Becomes less than one. When we compare smallest values of tiled bitmap algorithm with our approach, (lg(Iv))/Iv yields even smaller value than tiled bitmap. So we can state that our approach is having smallest cost of all algorithms.

This quantification of cost also reflects the space complexity of the forensic algorithms since each of thecontacts with the external notarization service correspondsto a hash value (of chains) which must be initially computed (and recomputed for comparison during validation) and maintained within the system. None of algorithms in Table 2 require extra space over the collection of hash values themselves.

4. CONCLUSION

Forensic analysis commences when a crime has been detected, in this case the tampering of a database. Such analysis endeavors to ascertain when the tampering occurred, and what data were altered. The present paper expands upon that work by presenting the Tiled Bitmap Algorithm, which is cheaper and more powerful than prior algorithms. This algorithm employs a logarithmic numberof hash chains within each tile to narrow down the whenand what. Checking the hash chain values produces abinary number; it is the task of the algorithm to computethe pre image of bitwise We also note that previous algorithms do not handle multiple corruption events well, whereas the Tiled Bitmap Algorithm can independently analyze corruption events occurring both in different tilesand multiple corruption events occurring within a single tile.

5. REFERENCES

[1] K.E. Pavlou and R.T. Snodgrass.2010"The tiled bitmap forensic analysis algorithm", IEEE transaction on knowledge and data engineering,Vol. 22, pp no.590-601, April 2010

[2] CSI/FBI. 2009. "Tenth Annual Computer Crime and Security Survey, http://www.cpppe.um.edu/Bookstore/ Documents/2005CSISurvey.pdf, 2009.

[3] M. Malmgren, 2009. "An Infrastructure for Database Tamper Detection and Forensic Analysis", honors thesis, University of Arizona, available at http://www.cs.arizona.edu/projects/tau/tbdb/MelindaMal mgrenThesis.pdf, 2009.

[4] U.S. Dept. of Health & Human Services. 2009. "The Health Insurance Portability and Accountability Act (HIPAA)," available at http://www.cms.hhsGov/HIPAAGenInfo/, 2009.

[5] J. Mena, Butterworth Heinemann. 2003."Investigative Data Mining for Security and Criminal Detection. 2003.

[6] M.T. Goodrich, M.J. Atallahand, and R. Tamassia, 2005. "Indexing Information for Data Forensics", Proc. Conf. Applied Cryptography and Network Security, pp. 206-221, 2005.

[7] R.T. Snodgrass, S.S. Yao, and C. Collberg, 2004. "Tamper Detection in Audit Logs", Proc. Int'l Conf. Very Large Databases, pages 504-515, Sept.2004.

[8] K.E. Pavlou and R.T. Snodgrass. 2006. "Forensic Analysis of Database Tampering", Proc. ACM SIGMOD Int'l Conf. Management of Data, pages 109-120, June 2006.

[9] K.E. Pavlou and R.T. Snodgrass. 2008."Forensic Analysis of Database Tampering", ACM Trans. Database Systems, vol. 33, no. 4, pages 1-47,Nov. 2008.

[10] Microsoft SQL Server 2000 [Online] Available: http://www.quackit.com/sql_server/sql_servr_2000/tutorial/about_sql_server.cfm

[11] Yamanishi, K. and Maruyama, Y., "Dynamic syslog mining for network failure monitoring," In KDD'05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pp. 499–508, New York, NY, USA. ACM, 2005.

[12] Facca, F. M. and Lanzi, P. L., "Mining interesting knowledge from weblogs: a survey," Data Knowledge Engineering, vol. 53(3), pp. 225–241, 2005.

[13] Li, D. H., Laurent, A., and Poncelet, P., "Mining unexpected web usage behaviors," In ICDM, pages 283–297, 2008.

[14] Harmeet Kaur Khanuja and Dr. D. S. Adane, 2011. "Database Security Threats and challenges in Database Forensic: A survey", Proceedings of 2011 International Conference on Advancements in Information Technology (AIT 2011), available at http://www.ipcsit.com/vol20/33-ICAIT2011- A4072.pdf, 2011.

[15] Kyriacos Pavlou, 2011. "Database Forensics in the Service of Information Accountability", Available at http://www.cs.arizona.edu/projects/tau/dragoon/

[16] Paul M. Wright, 2005. "Oracle Database Forensics using LogMiner", June 2004 Conference, SANS Institute, pages 1-39, 2005.

[17] Geoff H. Fellows, 2005. "The joys of complexity and the deleted file", Digital Investigation (Elsevier), Vol. 2, pages 89-93, February 2005.

[18] Garfinkel Simson, 2006. "Forensic feature extraction and cross-drive analysis", Digit Investigation (Elsevier), http://www.dfrws.org/2006/proceedings/10-Garfinkel.pdf, pages 71-81, August, 2006.

[19] Carrier Brian, 2005. "The Sleuth Kit & Autopsy: forensics tools for Linux and other Unixes", http://www.sleuthkit.org, 2005.

[20] Ewa Huebner, Derek Bem, Cheong Kai Wee, 2006. "Data hiding in the NTFS file system", Digital Investigation (Elsevier), pages 211-226, March 2006.

[21] D. Litchfield, 2007. "Oracle Forensics Redo Logs," NGS Software Insight Security Research (NI- SR), Next Generation Security Software Ltd., Sutton, 2007.

[22] D. Litchfield, 2007. "Oracle Forensics Part 2: Locating Dropped Objects," NGS Software Insight Security Research (NISR), Next Generation Security Software Ltd., Sutton, 2007.

[23] D. Litchfield, 2007. "Oracle Forensics Part 3: Isolating Evidence of Attacks against the Authentication Mechanism," NGSSoftware Insight Security Research (NISR), Next Generation Security Software Ltd., Sutton, 2007.

[24] D. Litchfield, 2007. "Oracle Forensics Part 4: Live NGSSSoftware Insight Security Research (NISR), Next Generation Security Software Ltd., Sutton, 2007.

[25] D. Litchfield, 2007. "Oracle Forensics Part 5: Finding of Data Theft in the Absence of Auditing," NGSSoftware Insight Security Research (NISR), Next Generation Security Software Ltd., Sutton, 2007.

[26] D. Litchfield 2007. "Oracle Forensics Part 6: Ex Segments, Flashback and the Oracle Recycle Bin," NGS Software Insight Security Research (NISR), Next Generation Security Software Ltd., Sutton, 2007.

[27] D. Litchfield, 2008. "Oracle Forensics Part 7: Using System Change Number in Forensic Investigations," NGSSoftware Insight Security Research (NISR), Next Generation Security Software Ltd., Sutton, 2008.