# Ensuring Data Provenance with Package Watermarking

Muhammad Umer Sarwar*, Muhammad Kashif Hanif†, Ramzan Talib‡, and Muhammad Asad Abbas§
Department of Computer Science,
Government College University, Faisalabad, Pakistan

***Abstract*—The last decade has shown tremendous growth data production from different sectors, e.g., biology, financial markets, scientific computing, business processes, Internet of Things. The "Data is New Oil" has become a proverb in academic and corporate circles. Accordingly, tracing, recording origin and deriving data called data provenance has gained tremendous traction across board. Privacy and security of data are major challenges to provenance management. This can be tackled using watermarking. The downside of majority of existing watermarking techniques is data distortion. In this work, we purpose a novel approach called package watermarking that addresses the data capacity, usability, robustness, security, distortion, verifiability, and detectability issues in data provenance.**

***Keywords*—Data security; Provenance; Watermarking; Tempering; Cryptography; Encryption; Decryption**

## I. Introduction

In the era of technology, the volume and complexity of data produced is increasing exponentially. The growth of data poses the concerns about data integrity and intellectual property protection. Tracking origin and history of data is an important task. Provenance or lineage is the mechanism to identify the ownership and derivation of data. Data trustworthiness can be evaluated using data provenance. This approach facilitates the detection of any type(s) of changes in the data and help to fix the responsibility for that change. It plays a vital role for the management, authenticity, integrity and trustworthiness of scientific data, relational database, semantic web, artwork, and digital objects [1]–[3]. Unstructured/semi-structured data, transparency of distribution, and interoperability of storage formats are major challenges to data provenance [4], [5].

Provenance systems can be classified into database-oriented, service-oriented, and miscellaneous categories [6], [7]. Researchers have extensively studied different provenance techniques in various applications domains with different properties (granularity, representation schemes, backend, overhead etc.) [6].

Researchers have also used watermarking techniques to ensure integrity and security of data [8], [9]. A watermark is embedded into the data for temper detection, ownership proof and traitor tracing [10]. The watermark must be invisible and difficult to remove [11]. There exist various watermarking techniques for data provenance including digital [12], fragile [13], visible [14], invisible [15], novel [16]. These techniques address data capacity, usability, robustness, security, distortion, verifiability and detectability issues. However, they are unable to cope with these issues to optimum level.

This paper presents a distortion free watermarking technique called package watermarking. This technique has security, integrity, verifiability, detectability, usability, usability, and robustness features. The rest of the article is divided into different sections. Section II presents related work. Section III describes the proposed methodology. Section IV presents scenario with the help of case study. The results are discussed in section V. Finally, section VI gives conclusion.

## II. Related Work

The growth of data poses the concerns about data integrity and intellectual property protection. Digital watermarking techniques have been employed for multimedia data. However, it was difficult to watermark relational data. Researchers have proposed different database watermarking techniques which can be categorized based on type of the watermark information, cover type, granularity level, verifiability, intent of marking, and distortion [10]. These techniques can be further characterized by data capacity, usability, robustness, security, and blindness [10].

Tiwari and Sharma studied various semi fragile watermarking algorithms using various image quality matrices, insertion and verification methods. However, issues of data capacity, usability and distortion are not addressed in semi fragile water marking [17]. Zhang et al proposed gray scale watermark pre-processing technique which greatly increases the robustness and capacity of the video watermarking for copyright protection. This technique maintains the good visual quality and almost the same bit rate. However, distortion from illegal attacks and verification at the granularity level (bit level) are major concerns [18].

Bartolini et al analyzed the performance of ST-DM watermarking in the presence of non-additive noise. They showed the gain attack plus additive Gaussian noise and the quantization attack affect the robustness and cause distortion in the ST-DM watermarking technique. This limits the effectiveness of this technique [19]. Noore proposed a semi-blind digital watermarking technique using the modified discrete cosine transformation. The results overcomes the attack issues but

can not eliminate distortion. This technique is not completely robust against the attacks [12].

Table I compares capacity, usability, robustness, security, distortion, and verifiability characteristics support in different studies of watermarking techniques. ⊗, ◔, ◑, ◕, and ● symbols represents the 0%, 25%, 50%, 75%, and 100% presence of issues, respectively. This symbol ○ represent "Not to be known" representation.

## III. Proposed Methodology

Databases have data confidentiality problem. We are proposing an encryption technique for data concealment or confidentiality in databases. Proposed technique employs symmetric key and stream cipher algorithm. It supports poly alphabetic substitution technique. Poly alphabetic substitution approach obtains cipher character by modular addition of plain text and key character (both should equal in length). Each cipher characters can substituted by several different characters.

Proposed algorithm has key generation, encryption, and decryption steps. In proposed technique, key will be generated between database user who will upload the file and the database service provider. It is the symmetric key algorithm so encryption and decryption key will remain same. Key will be user name and date of file upload. Next, uploaded file will be encrypted and then stored in database. Decryption is the reverse process of encryption it works in same way as encryption but in reverse. The proposed technique uses package to ensure data provenance in database systems. The package consists of encryption and decryption functions.

The features of proposed approach are:

- Proposed scheme is based on symmetric key algorithm that is much faster than asymmetric key algorithm.
- Key generation mechanism is very strong and unpredictable.
- Unique cryptographic key for each user.
- It follows poly alphabetic substitution method that replaces plain text character with multiple cipher characters.
- Frequency analysis and cryptanalysis is very difficult that makes our technique much secure.

## IV. Case Study

Every organization has its own organizational structure. Organizational structure determines how the roles, power and responsibilities are assigned, controlled, coordinated, and information flow between the different levels of management. Organizational structure can be centralized or decentralized depending on the organizations objectives and strategy.

The proposed data provenance approach is applied for information management system at Government College University, Faisalabad (Figure 1). The university follows the centralized organizational structure. The information management system has four departments, i.e., information, registration, accounts, and IT departments.

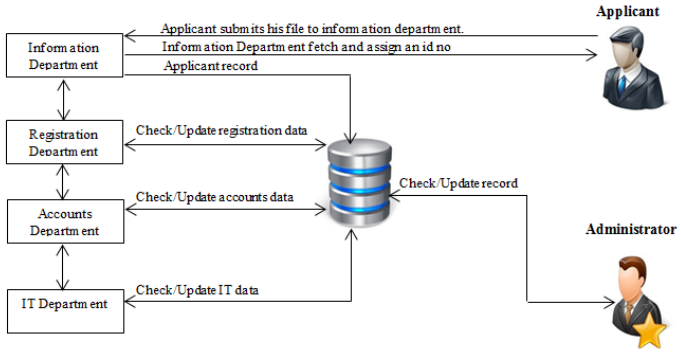



Fig. 1. Data flow in Government College University, Faisalabad.

Information department deals with different student queries. For example, if an applicant wants to take admission then he contacts information department to submit the admission application along with other documents. Information department will process the application and store in database by assigning a unique identifier to the applicant. Moreover, if an applicant require modification of his information then the department will forward request to other departments. Registration department will check the applicant record from database. For fresh students, registration department will assign the registration number. If student has already registration number then his data is verified from the database. In both cases, student data is updated in database and forwarded to the accounts department. Management of financial resources is important for any organization. In our case study, the account department is responsible to check, process, and store the financial data of the students from the database. Account department coordinates with IT and registration departments to handle queries. In current era of the technology, IT department is most important for an organization for smooth working. IT department administer and manage the database and computer network. IT department is responsible to answer different queries of other departments. Administrator acts like a super user who have administrative rights to add, update, and delete records. However, there exist no mechanism to check and track the tempering of data.

Data provenance can be used to check origin, transformation, and tempering of data. We have proposed information classification with respect to provenance at Government College University, Faisalabad (Figure 2). Applicant and information department can serve as origin of the data. Information, registration, accounts, and IT departments can modify the data. Administrator can modify the data and check the tempering. We have applied the package watermarking to ensure the provenance of the data.

## V. Results and Discussion

Organizations rely heavily on data generated by different business processes such customer relationship management, purchase management, inventory management. Tempering of data can affect the organization business. There are situations when data can be modified by illegally without knowledge of the users of data. It is difficult to find the tempering of

TABLE I. COMPARISON OF DIFFERENT WATERMARKING TECHNIQUES

| Author | Capacity | Usability | Robustness | Security | Distortion | Verifiability |
|---|---|---|---|---|---|---|
| Tiwari and Sharma [17] | ◔ | ◔ | ◔ | ◑ | ● | ◑ |
| Zhang et al [18] | ◑ | ◑ | ◕ | ◑ | ⊗ | ○ |
| Bartolini et al [19] | ◔ | ◔ | ◔ | ◔ | ● | ○ |
| Noore [12] | ◔ | ◔ | ◔ | ○ | ● | ◔ |

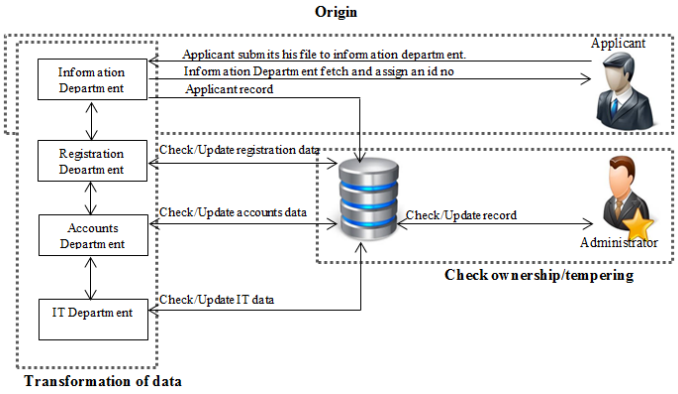
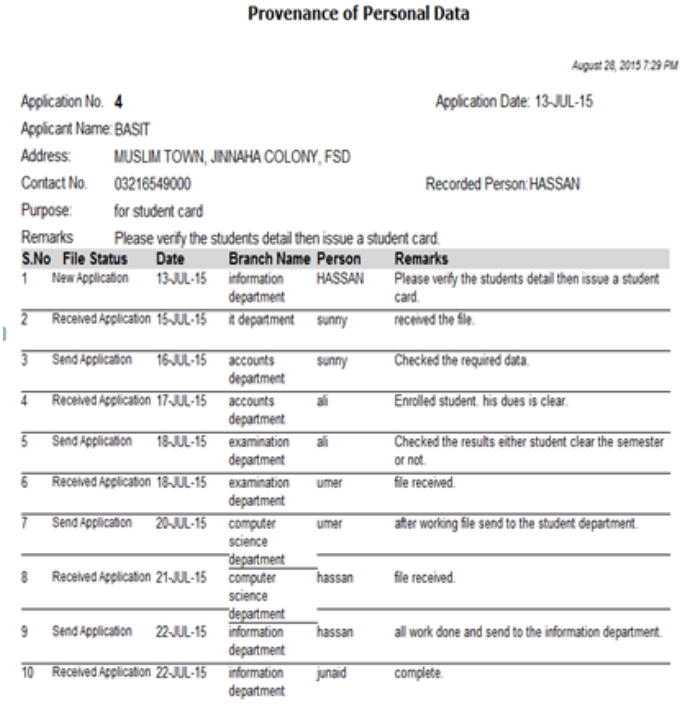



Fig. 2. Information classification with respect to provenance at Government College University, Faisalabad.



Provenance of Personal Data

August 28, 2015 7:29 PM

Application No. 4 Application Date: 13-JUL-15
Applicant Name: BASIT
Address: MUSLIM TOWN, JINNAHA COLONY, FSD
Contact No. 03216549000 Recorded Person:HASSAN
Purpose: for student card
Remarks Please verify the students detail then issue a student card.

| S.No | File Status | Date | Branch Name | Person | Remarks |
|---|---|---|---|---|---|
| 1 | New Application | 13-JUL-15 | information department | HASSAN | Please verify the students detail then issue a student card. |
| 2 | Received Application | 15-JUL-15 | it department | sunny | received the file. |
| 3 | Send Application | 16-JUL-15 | accounts department | sunny | Checked the required data. |
| 4 | Received Application | 17-JUL-15 | accounts department | ali | Enrolled student. his dues is clear. |
| 5 | Send Application | 18-JUL-15 | examination department | ali | Checked the results either student clear the semester or not. |
| 6 | Received Application | 18-JUL-15 | examination department | umer | file received. |
| 7 | Send Application | 20-JUL-15 | computer science department | umer | after working file send to the student department. |
| 8 | Received Application | 21-JUL-15 | computer science department | hassan | file received. |
| 9 | Send Application | 22-JUL-15 | information department | hassan | all work done and send to the information department. |
| 10 | Received Application | 22-JUL-15 | information department | junaid | complete. |

Fig. 3. View of data for authorized users after applying provenance.

data. The proposed approach will help for data trustworthiness. By implementing the proposed approach, organizations can be assured that data is not changed illegally. If someone has changed the data then it will be easy to track. The proposed approach will help organizations to secure data.

The proposed technique is applied to the case study presented in previous section. Figure 3 shows simple provenance of personal data. The results depicts a complete transformation of data and ownership of each department during tracking of data. This report can be only seen by the owner/administrator. Other person whether the user of the system or any intervener can never see this report.

After applying watermarking through encryption and decryption technique, we secured the provenance process as well as the data of the database. When any user other than administrator try to approach or access the data, he will not see the actual data (Figure 4). Figures 3 and 4 are same with same data, same application. However, Figure 4 is not in useable format since it does not any record. It is just a trash or nothing else for the intervener. It does not effect the data present in the database.

## VI. CONCLUSION

There are lots of benefits of using provenance such as show exact ownership, easily fulfill transformation of data, improved accessibility etc. However, there are yet practical problems in this technique that needs to be resolved. Data confidentiality is one of the major problem. Many researchers contributed their efforts to minimize the data security issue in this domain with different solutions. Cryptography is most widely used technique for data concealment in database domain. In this research work we proposed a new watermarking technique for data concealment in provenance environment in database. Security analysis of proposed approach proves that our approach is much secure against brute force attack, cryptanalysis, pattern prediction and frequency analysis. This technology needs the serious attention of the research community to gain the trust and confidence of databases users. In future, the proposed methodology will be applied to different types of data and environments.

## REFERENCES

[1] S. Haas, S. Wohlgemuth, I. Echizen, N. Sonehara, and G. Müller, "Aspects of privacy for electronic health records," *International journal of medical informatics*, vol. 80, no. 2, pp. e26–e31, 2011.

[2] L. Di, P. Yue, H. K. Ramapriyan, and R. L. King, "Geoscience data provenance: An overview," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 11, pp. 5065–5072, 2013.


Fig. 4. View of data for un-authorized users after applying proposed provenance technique.

[3] D. De Oliveira, V. Silva, and M. Mattoso, "How much domain data should be in provenance databases?" in *Proceeding of the 7th USENIX Workshop on the Theory and Practice of Provenance (TaPP 15)*, 2015.

[4] B. Glavic, "Big data provenance: Challenges and implications for benchmarking," in *Specifying big data benchmarks*. Springer, 2014, pp. 72–80.

[5] V. N. Gudivada, R. A. Baeza-Yates, and V. V. Raghavan, "Big data: Promises and problems." *IEEE Computer*, vol. 48, no. 3, pp. 20–23, 2015.

[6] Y. L. Simmhan, B. Plale, and D. Gannon, "A survey of data provenance techniques," Computer Science Department, Indiana University, Tech. Rep. 612, 2005.

[7] H. L. Dalpra, G. C. B. Costa, T. F. Sirqueira, R. M. Braga, F. Campos, C. M. L. Werner, and J. M. N. David, "Using ontology and data provenance to improve software processes." in *ONTOBRAS*, 2015.

[8] K. Ramya, R. C. Devi, M. Revathi, and P. Annapandi, "Sensor data hiding based on image watermarking using interpolation technique over inter-packet delays." *Applied Mechanics & Materials*, no. 573, 2014.

[9] M. D. Swanson, M. Kobayashi, and A. H. Tewfik, "Multimedia data-embedding and watermarking technologies," *Proceedings of the IEEE*, vol. 86, no. 6, pp. 1064–1087, 1998.

[10] R. Halder, S. Pal, and A. Cortesi, "Watermarking techniques for relational databases: Survey, classification and comparison." *J. UCS*, vol. 16, no. 21, pp. 3164–3190, 2010.

[11] M. Barni, F. Bartolini, A. De Rosa, and A. Piva, "A new decoder for the optimum recovery of nonadditive watermarks," *IEEE Transactions on Image Processing*, vol. 10, no. 5, pp. 755–766, 2001.

[12] A. Noore, "An improved digital watermarking technique for protecting jpeg images," in *2003 IEEE International Conference on Consumer Electronics, 2003. ICCE.*, 2003, pp. 222–223.

[13] M. Schäler, S. Schulze, R. Merkel, G. Saake, and J. Dittmann, "Reliable provenance information for multimedia data using invertible fragile watermarks," in *British National Conference on Databases*. Springer, 2011, pp. 3–17.

[14] S.-K. Yip, O. C. Au, C.-W. Ho, and H.-M. Wong, "Lossless visible watermarking," in *2006 IEEE International Conference on Multimedia and Expo*. IEEE, 2006, pp. 853–856.

[15] S. P. Mohanty and B. K. Bhargava, "Invisible watermarking based on creation and robust insertion-extraction of image adaptive watermarks," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 5, no. 2, pp. 1–22, 2008.

[16] L. Zhang, Y. Zhu, and L.-M. Po, "A novel watermarking scheme with compensation in bit-stream domain for h. 264/avc," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 1758–1761.

[17] A. Tiwari and M. Sharma, "Semifragile watermarking schemes for image authentication-a survey," *International Journal of Computer Network and Information Security*, vol. 4, no. 2, pp. 43–49, 2012.

[18] J. Zhang, A. T. Ho, G. Qiu, and P. Marziliano, "Robust video watermarking of h. 264/avc," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 54, no. 2, pp. 205–209, 2007.

[19] F. Bartolini, M. Barni, and A. Piva, "Performance analysis of st-dm watermarking in presence of nonadditive attacks," *IEEE Transactions on Signal Processing*, vol. 52, no. 10, pp. 2965–2974, 2004.