IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

# Survey on Data Leakage and Data Misusuability Issues

**O Sandeep Kumar[*1], B.Sunitha Devi[2]**

[*1]PG Student, Department of Computer Science & Engineering, CMR Institute of Technology, Hyderabad, India

[2]Associate Professor, Department of Computer Science & Engineering, CMR Institute of Technology, Hyderabad, India

odelasandeep538@gmail.com

## Abstract

Computer systems and the information that they generate, practice, transmit, and hoard has become obligatory to the contemporary endeavor. In today's on-demand, always connected, data-driven world and more than ever in light of the renovation of entire national economies from manufacturing-based paradigms to knowledge based ones—many organizations rightly count their information systems among their most important assets. Organizations often use these IT systems to store and process vast quantities of sensitive data, which, if disclosed, could be potentially damaging to an organization. Detecting and preventing data leakage and data misuse poses a serious challenge for organizations, especially when dealing with insiders with legitimate permissions to access the organization's systems and its critical data. This paper presents a survey on various data access control mechanisms used in the modern information world.

**Keywords**: Access Control, Data Confidentiality, Data Leakage, Data Misusability, Security.

## Introduction

In today's technological world the very important asset for every company is the Data. Some of these data are worth millions of dollars and organizations takes at most care in controlling the access to these data, with respect to both internal users, within the organization and external users outside the organization. Sensitive information such as customer or patient data and business secrets constitute the main assets of an organization. Such information is essential for the organization's employees, subcontractors, or partners to perform their tasks. Conversely, limiting access to the information in the interests of preserving secrecy might damage their ability to implement the actions that can best serve the organization.

Security breaches are typically categorized as unauthorized data observation, incorrect data modification, and data unavailability. Unauthorized data observation results in the disclosure of information to users not entitled to gain access to such information. All organizations, ranging from commercial organizations to social organizations, in a variety of domains such as healthcare and homeland protection, may suffer heavy losses from both financial and human points of view as a consequence of unauthorized data observation. Incorrect modifications of data, either intentional or unintentional, result in an incorrect database state. Any use of incorrect data may result in heavy losses for the organization. When data is unavailable, information crucial for the proper functioning of the organization is not readily available when needed.

Thus, data leakage and data misuse detection mechanisms are essential in identifying malicious insiders. The task of detecting malicious insiders is very challenging as the methods of deception become more and more sophisticated. According to the 2010 Cyber Security Watch Survey [1] 26 percent of the cyber-security events, recorded in a 12-month period, were caused by insiders. These insiders were the most damaging with 43 percent of the respondents reporting that their organization suffered data loss. Of the attacks, 16 percent were caused by theft of sensitive data and 15 percent by exposure of confidential data.

Data security is also crucial when addressing issues related to privacy of data pertaining to individuals; companies and organizations managing such data need to provide strong guarantees about the confidentiality of these data in order to comply with legal regulations and policies [2]. Overall, data security has a central role in the larger context of information systems security. Therefore, the development of Database Management Systems (DBMS) with high-assurance security is a central research issue. This paper completely focuses on the varuous techniques to control data access. The section II talks about the basic terminology of data access system. Section III presents a survey on various data access control mechanism used in the present data industry. Section IV concludes the survey.

## Data Leakage and Data Publishing

### *Basic Definitions*

Sensitive data means any data within an organization that is vital to the organization's core business. Data leakage is defined as the accidental or unintentional distribution of private or sensitive data to an unauthorized party. Sensitive data in organizations and companies include intellectual property (IP), financial information, patient information, personal credit card data and other information depending on the business or other industry. Furthermore in many cases sensitive data are shared among stakeholders such as employees working from outside, business partners and customers.

Data protection is ensured by different components of a database management system (DBMS). In particular, an access control mechanism ensures data confidentiality. Whenever a subject tries to access a data object, the access control mechanism checks the rights of the user against a set of authorizations, stated usually by some security administrator. An authorization states whether a subject can perform a particular action on an object.

Data confidentiality is further enhanced by the use of encryption techniques, applied to data when being stored on secondary storage or transmitted on a network. Recently, the use of encryption techniques has gained a lot of interest in the context of outsourced data management; in such contexts, the main issue is how to perform operations, such as queries, on encrypted data [3. Data integrity is jointly ensured by the access control mechanism and by semantic integrity constraints.

### *Privacy Preserving Data Publishing*

In the *data collection* phase, the *data publisher* collects data from *record owners* (e.g., Alice and Bob). In the *data publishing* phase, the data publisher releases the collected data to a data miner or to the public, called the *data recipient*, who will then conduct data mining on the published data. In this survey, data mining has a broad sense, not necessarily restricted to pattern mining or model building. For example, a hospital collects data from patients and publishes the patient records to an external medical center. In this example, the hospital is the data publisher, patients are record owners, and the medical center is the data recipient. The data mining conducted at the medical center could be anything from a simple count of the number of men with diabetes to a sophisticated cluster analysis.

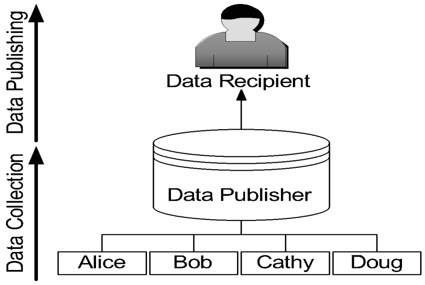
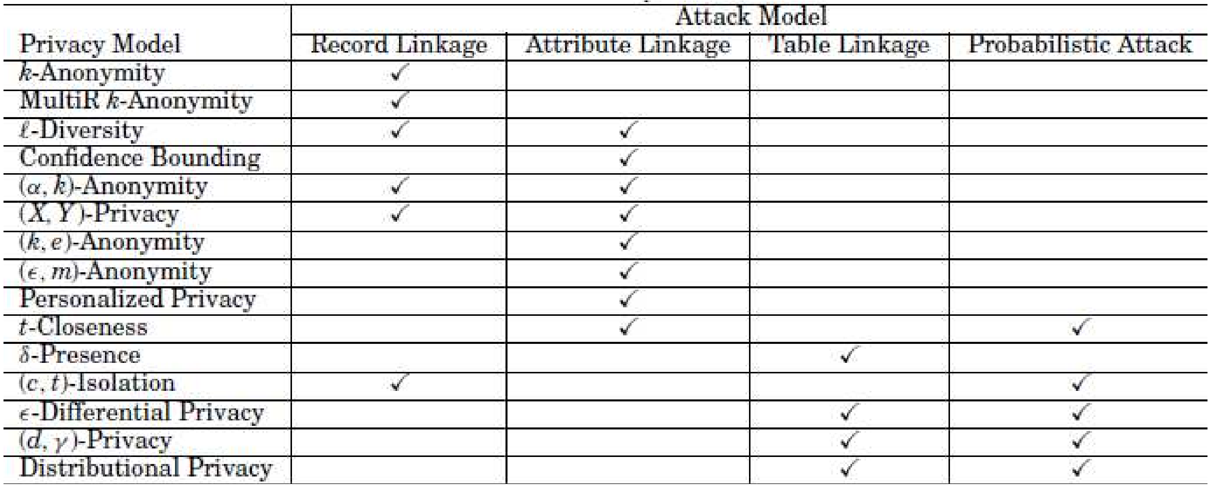


**Fig 1 Data collection and Publishing**

There are two models of data publishers [Gehrke 2006]. In the *untrusted* model, the data publisher is not trusted and may attempt to identify sensitive information from record owners. Various cryptographic solutions; anonymous communications ; and statistical methods were proposed to collect records anonymously from their owners without revealing the owners' identity. In the *trusted* model, the data publisher is trustworthy and record owners are willing to provide their personal information to the data publisher; however, the trust is not transitive to the data recipient. In this survey, we assume the trusted model of data publishers and consider privacy issues in the data publishing phase.

During the past decade, several measures in the field of Privacy-Preserving Data Publishing were introduced [8]. Examples of such measures are k-Anonymity [4], l-Diversity [5], and (α,k)-Anonymity. These measures attempt to estimate how easy it is to compromise an individual's privacy in a given publication, where publication refers to a table of data containing quasi-identifier attributes, sensitive attributes, and additional attributes. The main goal of these measures is to estimate the ability of an attacker to infer who are the individuals (also called victims) behind the quasi-identifier, and thus reveal sensitive attribute values (e.g., disease).

PPDP algorithms are useful when there is a need for exporting data (e.g., for research) while retaining the privacy of individuals in the published data set. It can also be used in a limited way for estimating the level of Misusability of data. The harder it is to identify who is the entity in a record, the lower the potential risk of a perpetrator maliciously exploiting that information. This approach, however, is not effective in other scenarios that assume a user has full access to the data. Sweeney [6] proposed the k-anonymity measure that indicates how hard it is to fully identify who is the entity in each record in a published table T, given a publicly available database (e.g., Yellow Pages). The measure determines that T satisfies k-anonymity if and only if each value of the quasi-identifier in T appears at least k times. A known disadvantage of k-anonymity is that it does not consider the diversity of the sensitive attribute value (also known as the common sensitive attribute problem). This approach is relevant only when exposing statistical information rather than individual records (e.g., for analytics or data mining tasks). However, in most cases, performing different tasks require exposing the individual records. TheM-score measure is mainly used for deriving the misuseability level of the individual records exposed to the user.. The following Table shows the privacy model and the attack model.



| Privacy Model | Attack Model<br>Record Linkage | <br>Attribute Linkage | <br>Table Linkage | <br>Probabilistic Attack |
|---|---|---|---|---|
| $k$-Anonymity | ✓ | | | |
| MultiR $k$-Anonymity | ✓ | | | |
| $\ell$-Diversity | ✓ | ✓ | | |
| Confidence Bounding | | ✓ | | |
| $(\alpha, k)$-Anonymity | ✓ | ✓ | | |
| $(X, Y)$-Privacy | ✓ | ✓ | | |
| $(k, e)$-Anonymity | | ✓ | | |
| $(\epsilon, m)$-Anonymity | | ✓ | | |
| Personalized Privacy | | ✓ | | |
| $t$-Closeness | | ✓ | | ✓ |
| $\delta$-Presence | | | ✓ | |
| $(c, t)$-Isolation | ✓ | | | ✓ |
| $\epsilon$-Differential Privacy | | | ✓ | ✓ |
| $(d, \gamma)$-Privacy | | | ✓ | ✓ |
| Distributional Privacy | | | ✓ | ✓ |

## Misusability Weight Calculation

Generally the users in the organization performs various actions on the data which may affect the sensitive information associated with the data. If this data leaks to outside the perimeter of the organization it introduced loss to the organization. To measure the level of damage a new concept is used now days. That's termed as Misusability weight. This Misusability weight measures the amount of damage happened to the organization. A neat and clear procedure is required to calculate the Misusability weight. .

***Dimesnsions of Misusability***

The Misusability weight called as M-Score mainly depends on the domain of the data and sensitiveness of the data. The same data will have different M-Score in different domains. There are four dimensions of Misusability.

***Number of entities*** - This is the data size with respect to the different entities that appear in the data. Having data about more entities obviously increase the potential damage as a result of a misuse of this data.

***Anonymity level*** - While the number of different entities in the data can increase the misuseability weight, the anonymity level of the data can decrease it. The anonymity level is regarded as the effort that is required in order to fully identify a specific entity in the data.

***Number of properties*** - Data can include a variety of details, or properties, on each entity (e.g., employee salary or patient disease). Since each additional property can increase the damage as a result of a misuse, the number of different properties (i.e., amount of information on each entity) should affect the misuseability weight.

***Values of properties***- The property value of an entity can greatly affect the misuseability level of the data. For example, a patient record with disease property equals to HIV should probably be more sensitive than a record concerning patient with a simple flu.

***Dimesnsions of Misusability***

The basic building block of this M-Score is table. The Table T(A1; . . .;An) is a set of r records. Each record is a tuple of n values. The value i of a record, is a value from a closed set of values defined by Ai, the i's Attribute of T. Therefore, we can define Ai either as the name of the column i of T, or as a domain of values. The three non intersecting parameters are

1. *Quasi Identifier Attributes* are attributes that can be linked, possibly using an external data source, to reveal a specific entity that the specific information is about. In addition, any subset of the quasi-identifiers
2. *Sensitive Attributes* are attributes that are used to evaluate the risk derived from exposing the data.
3. Other Attributes.

From these attributes a sensitive score function f is calculated which is very important for the calculation of M-Score.

Sensitive Score Function -

The sensitivity score function f : C X $S_j$ -> [0; 1] assigns a sensitivity score to each possible value x of Sj, according to the specific context c 2 C in which the table was exposed. For each record r, we denote the value $X_r$of $S_j$ as $S_j$ [$x_r$].

***M-Score Calculation***

The M-score incorporates three main factors.

1. Quality of data—the importance of the information.

2. Quantity of data—how much information is exposed.

3. The Distinguishing Factor (DF)—given the quasiidentifiers,

The amount of efforts required in order to discover the specific entities that the table refers to. The calculation of the raw record score of record i ($RRS_i$), is based on the sensitive attributes of the table, their value in this record, and the table context. For a record i, $RRS_i$ will be the sum of all the sensitive values score in that record, with a maximum of 1.

Given a table with r records, the table's M-score is calculated as follows:

$$MScore = r^{1/x} \times RS = r^{1/x} \times \max_{0 \leq i \leq r} \left( \frac{RRS_i}{D_i} \right),$$

where r is the number of records in the table, x is a given parameter and RS is the final Record Score

## Conclusion

Written with the objective to shed light on existing data publishing techniques and the possible data leakage ways in an organization. If possible sources are analyzed then the protection will be very easy. The Protection is by assigning a weight for each sensitive data called as M-Score. The mechanism used to calculate the M-Score is also presented in this paper.

## References

[1] 2010 CyberSecurity Watch Survey, http://www.cert.org/ archive/pdf/ecrimesummary10.pdf, 2012.

[2] A. Kamra, E. Terzi, and E. Bertino, "Detecting Anomalous Access Patterns in Relational Databases," Int'l J. Very Large Databases, vol. 17, no. 5, pp. 1063-1077, 2008.

[3] S. Mathew, M. Petropoulos, H.Q. Ngo, and S. Upadhyaya, "Data- Centric Approach to Insider Attack Detection in Database Systems," Proc. 13th Conf. Recent Advances in Intrusion Detection, 2010.

[4] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," Int'l J. Uncertainty, Fuzziness and Knowledge Based Systems, vol. 10, no. 5, pp. 571-588, 2002.

[5] A. Machanavajjhala et al., "L-Diversity: Privacy Beyond K-Anonymity," ACM Trans. Knowledge Discovery from Data, vol. 1, no.1, article 1, 2007.

[6] R.C. Wong, L. Jiuyong, A.W. Fu, and W. Ke, "(_,k)-Anonymity: An Enhanced k-Anonymity Model for Privacy-Preserving Data Publishing," Proc. 12th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 2006.

[7] E. Celikel et al., "A Risk Management Approach to RBAC," Risk and Decision Analysis, vol. 1, no. 2, pp. 21-33, 2009.

[8] B. Carminati, E. Ferrari, J. Cao, and K. Lee Tan, "A Framework to Enforce Access Control over Data Streams," ACM Trans. Information