International Journal of Information and Computation Technology.
ISSN 0974-2239 Volume 3, Number 4 (2013), pp. 341-348
© International Research Publications House
http://www. irphouse.com /ijict.htm

Multiple Distinguishing Factors Based Measure for Data Misuseability Detection

G. Manogna[1] and K.F.Bharathi[2]

1Department of Computer Science, Engineering, JNTUA College of Engineering, Anantapur (Autonomous), India
[2]M.Tech., Department of Computer Science, and Engineering, JNTUA College of Engineering.

Abstract

Data Anantapur (Autonomous), India. is valuable for any business. The power of organization lies in the data it maintains. Preserving such data and secur ing it has to be given paramount importance. Misuse of such data is a challenge to organizations. The data in an or ganization is used by the insiders. It does mean that the data is exposed to internal users of organizations for performing legal and accepted activities. To prevent misuse the misusability of the data has to be estimated before exposing it to the insiders. Recently Harel et al. presented a novel concept known as "Misusability Weight" which is used to estimate the risk level of data when it is exposed to insiders. According to this concept, a score is associated with data which will reflect the sensitivity level of data being exposed. This will help in securing data from misusability. In this paper we extend the method followed by Harel et al. to support more than one distinguishing factor and sensitivity of combinations of sensitive values. We also built a prototype application that demonstrates the proof of concept. The empirical results revealed that the proposed extension to misusability measure has shown improvement in securing data from misusability.

Keywords: Data leakage, misusability weight, data mining.

1. Introduction

For any organization data is an asset. In all domains such as health care, banking, education, insurance and so on data plays an important role in doing business. The data is required to perform various kinds of transactions and also update it as transactions going on. The data access to such valuable data is limited through role based access control or any other means. However, when data access is provided to insiders of organizations, the insiders may misuse it. Assuming that there is possibility of misusability, it is essential to have a measure to know how sensitive the data exposed is. To overcome this problem certain mechanisms to prevent data misuse and data leakage is essential. However, finding the people who commit such malicious activities on the accessed data is very challenging for many reasons. Moreover there are incidents in the history that proved that insiders misuse data for some reasons including monetary gains. According a survey [1] around 26 percent of misuse events occurred only due to malicious insiders. Preventing the insider attacks is difficult when compared to preventing external attacks. When a research was carried out 43 percent people respondents said that data loss in their organization was caused by insiders. That is done due to theft of sensitive data and also exposure to confidential data.

Recently Harel et al. [30] proposed a novel concept known as “Misusability measure” to know the misusability score of the data being exposed to insiders. This will reduce the misuse of leakage of data which is made available to the insiders. User behavior profiles are used to devise plans to mitigate such fraud. When profiles are observed, it can be understood that the normal user’s behavior is obviously different from that of the malicious user who has malicious intentions. User behavior can be analyzed using SQL commands and other features provided by the query languages [2]. In [3] also another approach is presented. However, different sensitive levels are not used by the existing solutions. However, this feature gives great opportunity to know misusability positively. There are other data measures for ensuring security. They include K-Anonymity [6], and l-Diversity [5]. These techniques are mainly for privacy preserving and they are not relevant when user is to be given free access. For this reason the misuability measure presented in [3] is used to know the sensitivity score on datasets to reduce the harm level of valuable business data from being accessed by the insiders illegally. The following are the usages of the misusability measure presented in [30].

- Anomaly can be detected by studying the difference between the normal user’s behavior and the behavior of malicious users.
- Sensitivity concept is used to improve the system that can best prevent data from being misused.
- Dynamic Misusablity – Based Access Control (DMBAC) is used to have controlled access to sensitive data.
- Misusability of data is reduced.

In this paper we extend the concept used in [30] to further improve the misusability measure. We support multiple publications with more than one distinguishing factor and sensitivity of combinations of sensitive values. The remainder of this paper is structured as follows. Section II reviews literature. Section III presents the extended misusability weight concept and M-score measure. Section IV presents extended MScore measure. Section V presents prototype implementation. Section VI presents experimental results while section VII concludes the paper.

2. Related Work

Data leakage problem has been around for many years. When data is exposed to insiders in an organization, they may tend to misuse the data. Data leakage also takes many forms. There are many existing methods to combat this problem. These methods can be classified into two categories. The first category is based on data centric while the second is based on syntax centric. The latter is based on SQL and its syntax. The SQL supports four kinds of commands such as DML, DCL, TCL, and DCL. These commands are built based on certain syntax. Every operation is carried out using a query. It does mean that application can interact with backend only by issuing SQL queries such as SELECT, INSERT, UPDATE, DELETE and so on[2]. A risk management system was proposed in [7]. This model was able to find the risk level posed by specified user. Another model on syntax-centric method is using access control based on data streams [8].

On the other hand, the data centric approach gives importance to what user tries rather than what is expressed by user. This model results in more direct reflection of misuse case. Query's expression syntax is also presented in [3] in data centric approach. Another model known as S-Vector proved to be better than other approaches [2]. Another data centric approach presented in [8] uses dependency graphs to predict sensitive information. An insider predication model was proposed in [10]. It computes taxonomy pertaining to insider threats and evaluates potential threat. Group Based Access Control (GBAC) was proposed in [11] for preventing insider threats.

With respect to privacy preserving data publishing many techniques came into existence [12]. They are known as Anonymity [6], K-Anonymity [4], and l-Diversity [5]. These methods are meant for protecting data privacy while publishing it. These are useful when data has to be exposed to public for data mining or other purposes. Differential privacy is another research related to our topic. Its goal is to ensure privacy preserving of data [13], [14]. These schemes have limitations. To overcome the drawbacks Misusability measure was introduced by Harel et al. [30] to find the given data is misusable by insiders or not if yes to which extent.

3. M-score Measure

The misusability measure presented by Harel et al. [30] is used in this paper for enhancement. Therefore more details can be found in [30] on various details associated with M-Score. However, we have built a framework to visualize the process of measuring M-Score for the given dataset. The framework is as shown in fig. 1.

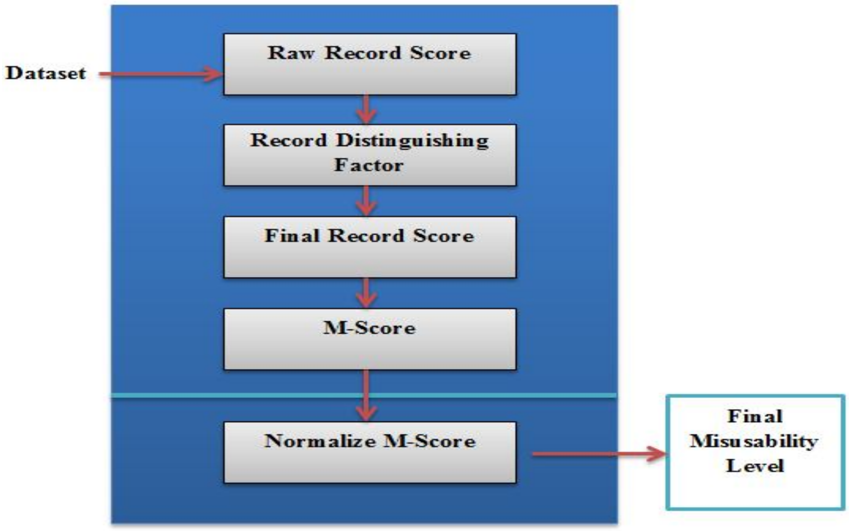


Fig. 1: Architecture for measuring M-Score (based on the taken from [30]).

As can be seen in fig. 1 there are many phases in finding final misusability level of given dataset which is supposed to be exposed to insiders of an organization. The final misusability level helps organizations to make well informed decisions to expose the data as it is or to protect it further. First of all the given dataset is given to a module which computes raw record score. Sensitivity score function is used to computer raw record score as follows.

$$RRS_i = \min\left(1, \sum_{S_j \in T} f(c, S_j[x_i])\right).$$

Afterwards the record distinguishing factor (DF) is computed. The DF is the measure to know what extent the quasi identifiers can reveal the identity of a record. The final Record Score is computed as follows.

$$RS = \max_{0 \le i \le r}(RS_i) = \max_{0 \le i \le r}\left(\frac{RRS_i}{D_i}\right).$$

Once record score is computed, the values are substituted in the formula of M-score. The computation of M-score is computed as follows.

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

The problem with MScore is that the value is unbounded. Therefore to make it bound between 0.0 and 1.0, it has to be normalized. The normalization can help users understand the meaning of MScore easily. The more the MScore value the more the misusability changes are.

4. Extended M-score Measure

The existing MScore measure does not support when DF value is more than 1. We have modified the measure to incorporate it. The corresponding framework we devised is presented in fig. 2.

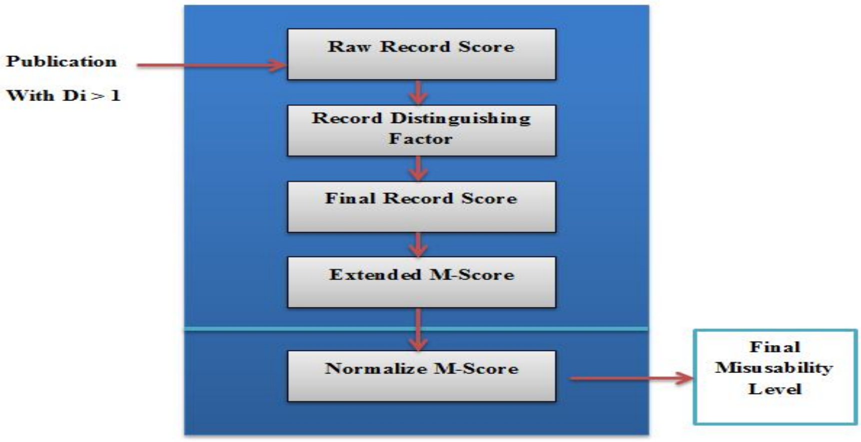
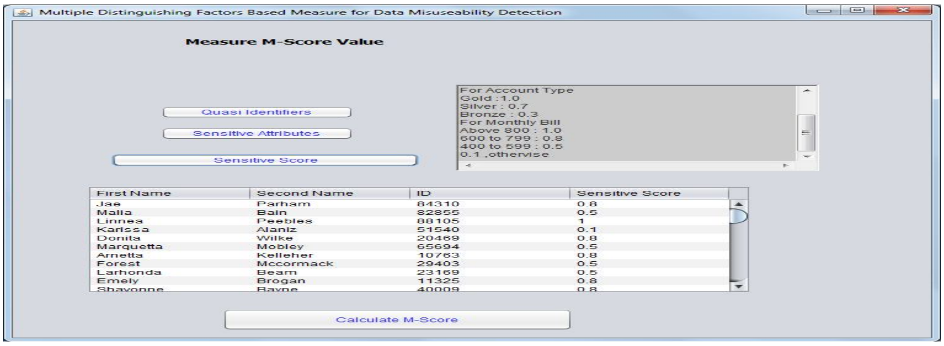


Fig. 2: Extended M-Score Measure.

As can be fig. 2 it is evident that the framework is capable of taking publications where DF value is higher than 1. It does mean that it supports multiple distinguishing factors to measure MScore. Finally result is the value that can be understood by end users. The final value is between 0.0 and 1.0 reflecting the level of misusability associated with given publication.

5. Prototype Implementation

We built a prototype application in order to demonstrate the efficiency of the extended M-Score proposed by us. The environment used is a PC with 4GB RAM, Core 2 dual processor running Windows 7 operating system. Java platform is used to build the application with user friendly interface.

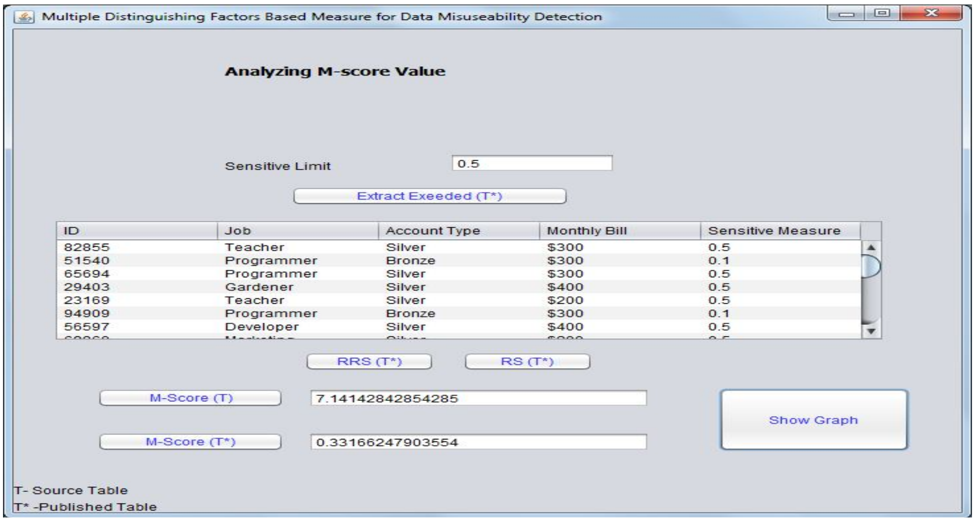


For Account Type
Gold :1.0
Silver : 0.7
Bronze : 0.3
For Monthly Bill
Above 800 : 1.0
600 to 799 : 0.8
400 to 599 : 0.5
0.1 ,othervise

First Name	Second Name	ID	Sensitive Score
Jae	Parham	84310	0.8
Malia	Bain	82855	0.5
Linnea	Peebles	88105	1
Karissa	Alaniz	51540	0.1
Donita	Wilke	20469	0.8
Marquetta	Mobley	65694	0.5
Arnetta	Kelleher	10763	0.8
Forest	Mccormack	29403	0.5
Larhonda	Beam	23169	0.5
Emely	Brogan	11325	0.8
Shavonne	Bayne	40009	0.8

Fig. 3: Sensitivity Score of Various Identifiers.

As can be seen in fig. 3, it is evident that the application loaded dataset and has provision for knowing the sensitivity of quasi identifiers, sensitive attributes and also to compute sensitive score. It also facilitates to choose calculation of final M-Score. On choosing “Calcualte M-Score”, the screen appears as shown in fig. 4.



Multiple Distinguishing Factors Based Measure for Data Misuseability Detection

Analyzing M-score Value

Sensitive Limit 0.5

ID	Job	Account Type	Monthly Bill	Sensitive Measure
82855	Teacher	Silver	$300	0.5
51540	Programmer	Bronze	$300	0.1
65694	Programmer	Silver	$300	0.5
29403	Gardener	Silver	$400	0.5
23169	Teacher	Silver	$200	0.5
94909	Programmer	Bronze	$300	0.1
56597	Developer	Silver	$400	0.5

M-Score (T) 7.14142842854285

M-Score (T*) 0.33166247903554

T- Source Table
T* -Published Table

Fig. 4: Final results.

As can be seen in fig. 3, for the given dataset and given sensitive limit the RRS and RS are computed. Then the M-Score is computed. As the resultant value of M-Score is not bounded, it has been normalized. The normalized value is shown in fig. 3. On choosing “Show Graph” a graph is generated to show the results of the existing M-Score as shown in fig. 5.

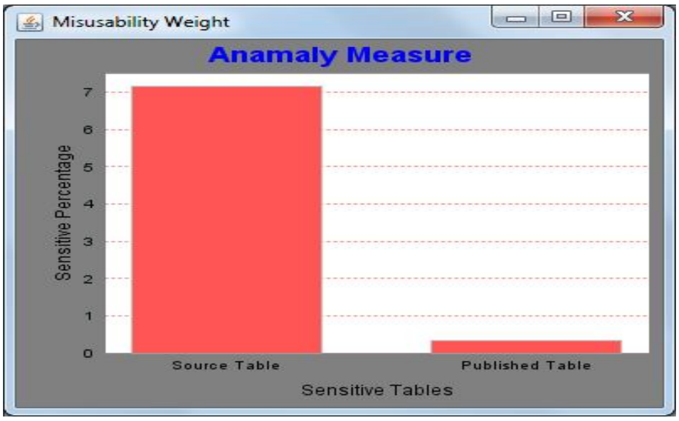


Fig. 5: Sensitive percentage of source table and published table.

As seen in fig. 5, it is evident that the published table has less sensitive percentage. The source table for which security is not applied is having higher sensitivity percentage.

6. Experimental Resutls

We focused on testing three datasets with proposed system which has improved M-Score measure that supports multiple publications with DF more than 1. The results are visualized as shown in fig. 6.

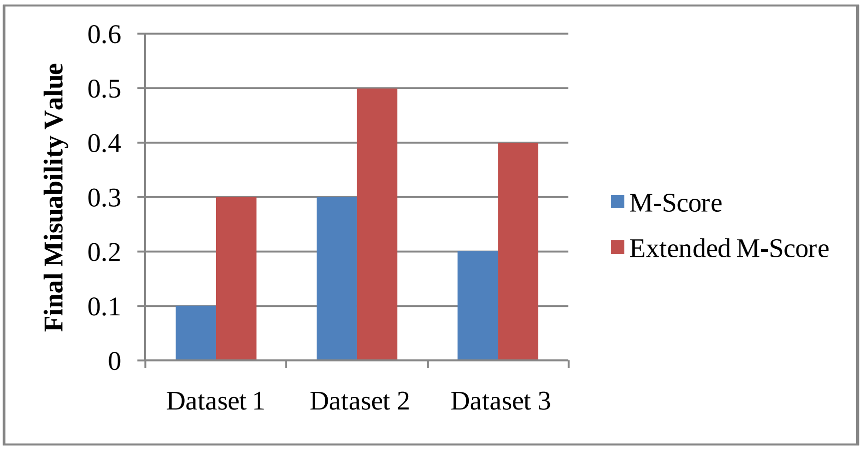


Fig. 6: Comparison of results.

As seen in fig. 6, the proposed system is capable of giving more accurate misusability measure as it is able to consider multiple DFs. For this reason in all the datasets, the results revealed that extended M-Score outperforms the existing M-Score.

7. Conclusion

In this paper we extend the concept by name "Misusability Measure" introduced by Harel [30] which can compute the misusability measure for given dataset. The result of this measure is between 0.0 and 1.0 which shows the level of misusability. It does mean that the data is exposed to insiders with given misusability possibilities. This measure is based on four dimensions. The dimensions include number of entities, anonymity level, number of properties, and values of properties. The misusability measure is also computed based on three main factors as well. They include quality of data, quantity of data, and the distinguishing factor. In this paper we extended the concept of Measurability weight to support multiple publications with more than one distinguishing factor and sensitivity of combinations of sensitive values. We also built a prototype application to demonstrate the proof of concept. The empirical results revealed that the proposed prototype is useful and can be used in real time applications to know the measure of misusability of data by insiders and make well informed decisions before making decisions pertaining to data exposure to insiders.

References

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

[2] Amir Harel, AsafShabtai, LiorRokach and Yuval Elovici, “M-Score: A Misuseability Weight Measure”. IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, VOL. 9, NO. 3, MAY/JUNE 2012.

[3] A. Kamra, E. Terzi, and E. Bertino, “Detecting Anomalous AccessPatterns in Relational Databases,” Int’l J. Very Large Databases,vol. 17, no. 5, pp. 1063-1077, 2008.

[4] S. Mathew, M. Petropoulos, H.Q. Ngo, and S. Upadhyaya, “Data-Centric Approach to Insider Attack Detection in DatabaseSystems,” Proc. 13th Conf. Recent Advances in Intrusion Detection,2010.

[5] R.C. Wong, L. Jiuyong, A.W. Fu, and W. Ke, “(_,k)-Anonymity:An Enhanced k-Anonymity Model for Privacy-Preserving DataPublishing,” Proc. 12th ACM SIGKDD Int’l Conf. KnowledgeDiscovery and Data Mining, 2006.

[6] A. Machanavajjhala et al., “L-Diversity: Privacy BeyondK-Anonymity,” ACM Trans. Knowledge Discovery from Data, vol.1, no.1, article 1, 2007.

[7] B. Carminati, E. Ferrari, J. Cao, and K. Lee Tan, “A Framework toEnforce Access Control over Data Streams,” ACM Trans. InformationSystems Security, vol. 13, no. 3, pp. 1-31, 2010.

[8] G.B. Magklaras and S.M. Furnell, “Insider Threat Prediction Tool:Evaluating the Probability of IT Misuse,” Computers and Security,vol. 21, no. 1, pp. 62-73, 2002.

[9] M. Bishop and C. Gates, “Defining the Insider Threat,” Proc. Ann.Workshop Cyber Security and Information Intelligence Research, pp. 1-3, 2008.

[10] C.M. Fung, K. Wang, R. Chen, and P.S. Yu, “Privacy-PreservingData Publishing: A Survey on Recent Developments,” ACMComputing Surveys, vol. 42, no. 4, pp. 1-53, 2010.

[11] L. Sweeney, “k-Anonymity: A Model for Protecting Privacy,” Int’lJ. Uncertainty, Fuzziness and Knowledge Based Systems, vol. 10, no. 5,pp. 571-588, 2002.

[12] A. Friedman and A. Schuster, “Data Mining with DifferentialPrivacy,” Proc. 16th ACM SIGKDD Int’l Conf. Knowledge Discoveryand Data Mining, pp. 493-502, 2010.

[13] C. Dwork, “Differential Privacy: A Survey of Results,” Proc. FifthInt’l Conf. Theory and Applications of Models of Computation, pp. 1-19,2008.