Comparison of Data Leakage Detection and Prevention Algorithms

Amey Sawant[1], Aghan Lawande[2], Dalton Fernandes[3], Poojan Vaigankar[4], Sarvesh Vani[5], Prof. Valerie Menezes[6], Prof. Kedar Sawant[7]

[1,2,3,4,5]*Department of Computer Engineering, Agnel Institute of Technology and Design, Goa, India*

[6,7]*Assistant Professor, Department of Computer Engineering, Agnel Institute of Technology and Design, Goa, India*

Abstract **- Data anonymization is process by which personal data is altered in such a way that a data subject can no longer be identified directly or indirectly and the process sometimes irreversible, either by the data controller alone or in collaboration with any other party. Data anonymization may by accident enable the transfer of information across a boundary, such as between two departments within an organization or between two agencies, while reducing the risk of inadvertent divulge, and in certain environments in a way that enables evaluation and analytical post-anonymization. In the context of medical data, anonymized data refers to data from which the patient cannot be identified by the recipient of the information but it's just enough to gain desired knowledge which will not affect privacy. The name, address, and postcode must be removed, together with any other information which, in cooperation with other data held by or disclosed to the recipient, could identify the patient. There has always been a risk that anonymized data may not stay anonymous for a long period of time. Pairing the anonymized dataset with other data, advanced techniques and raw power are some of the ways previously anonymous data sets have become de-anonymized. The data subjects are no longer anonymous.**

Index Terms **- Data Leakage, Data Anonymization, Prevention Algorithms.**

1.INTRODUCTION

Data leakage is defined as the accidental or unintentional loss of private or sensitive data to an unauthorized entity. It poses a serious issue for companies as the number of incidents and the cost to those experiencing them continue to increase. Data leakage is increased due to transmission of data through emails, messaging, website forms, and file transfers among others, are largely unregulated and unmonitored on their way to their destinations. When distributor's sensitive data that has been leaked by agents, it is not possible to identify the agents who leak the data only on the bases of a watermark technique which is used for keeping such data private. A model for calculating guilty agent probabilities in cases of data leakage is presented here. K-Anonymity algorithm is used to generalize and suppress sensitive data, so data sets will be hidden and third parties will not be able to view the original data sets.

A data leak happens when confidential data is accidentally exposed intentionally or by accident, on the Internet or any other form including lost hard drives or laptops. This means a cyber-criminal can gain unauthorized access to the sensitive data without effort and misuse it. The worst part is once a data exposure has happened it is extremely difficult to know whether the data was accessed or not. This means that your confidential data, trade secrets, source code, customer data, personal data and any type of confidential data stored on information systems could be exposed or used as part of corporate spying.

As technology advances, for many organizations and businesses data privacy and confidentiality has become the highest priority. To hide such data to manage confidentiality and also to utilize the information from the hidden data as much as possible, we use the anonymization process to anonymize datasets. Even if a normal or an anonymized dataset is leaked, we use the sample data and explicit data algorithms to trace the leak.

There are several different types of anonymizations algorithms which are used to anonymize data sets. For this project we'll be researching on sample data and explicit data request for data leakage detection on k anonymity and incognito algorithm for data leakage

prevention. Further after our research is completed, we'll be using test cases to compare both the algorithms based on prevention of data leakage. The utilization of sample data request and explicit data request algorithms will be a part of the research work and test cases will be utilized.

2. LITERATURE SURVEY

Security attacks that threaten the wellbeing of organizations are changing in various ways.[10] The most common being cyber-attacks those possess major threat to organization wellbeing as they are most come from outside sources, which makes them very difficult to identify. But there is possibility of security attacks coming from inside the organization. In today's world data being the most valuable asset a organization can have and every other organization want to have it. Insiders have access to data of that organization so it can be shared to attackers making insiders a big threat.By Rama Rajeswari Mulukutla & P. Poturaju states about various allocation strategies, data privacy, data leakage, leakage model, fake records. Using agent guilt model to calculate the probability that an agent has leaked the data that is guilty based on comparing the probability of the various agents having access to that sensitive data.

Agent makes two types of requests, called sample and explicit. Based on the request the Fake objects are added to the data list. Fake objects are objects generated by the distributor that are not in set T [1]. Where T stands for the original table of sensitive data. Furthermore, the allocation of fake objects has to be done is a way the data that is sent after adding fake objects is sensible enough. The number of fake objects to be allocated to a table so that it is sensible enough is determined by the data allocation algorithms (sample and explicit request algorithms).The working of these methods is explained in this research paper. ie. the sample data request algorithm and explicit data request algorithm.

Multiple anonymization algorithms have been proposed. However, given the large number of algorithms available and limited information regarding their performance, it is difficult to identify and select the most appropriate algorithm given a particular publishing scenario [2]. Which basically states that different tests on algorithms provides results that can be favourable for different purpose, so researchers can use the algorithm based on the results of the test that is optimistic for their research purpose. This is because it is not possible for one algorithm to give the best results for all possible test cases. And using a algorithm that is accurate for the desired research purpose will give a more precise result as compared to an optimal algorithm but doesn't favour the desired test scenario.In this research paper we compare the optimal k-anonymity and the Incognito algorithm based on three test cases that are information loss caused after data is anonymized, minimal distortion of data anonymized and calculation of execution time of the algorithms.

Top Down uses local-recoding based top-down approach, proposed this method which specialize and generalize the single data attribute one level down every iteration and unspecialized the data if it doesn't satisfy the k-anonymity condition [4].It uses tree like structure to achieve generalization hierarchy and each node is formed during the specialization of the data .We start with equivalence class of that particular attribute and apply top down algorithm .Once the generalization hierarchy has been formed the nodes are replaced with original data in the dataset which needs to be sent.

LeFevre, DeWitt and Ramakrishnan propose an efficient algorithm for computing k-minimal generalization, called Incognito, which takes advantage of a bottom-up aggregation along dimensional hierarchies and a priori aggregate computation [3]. This algorithm takes all the quasi-identifiers and creates generalization lattice out of the quasi attributes. We use modified breadth first search to determine which all nodes satisfy the k anonymity condition and those nodes from the lattice can be used for anonymizing the data depending on the need one can choose the level of privacy or utility of the data.

There are several different anonymization algorithms which are used to anonymize data sets. Water marking technique was used previously in data leakage detection algorithms but the disadvantage was that in this technique the original copy had to be modified. Another technique was Perturbation technique where in the data was mainly approximated to the nearby value of the original data i.e. the original data was modified. Compared to k- Anonymity algorithm and Incognito Algorithm these algorithms were quite inefficient.

3. PROPOSED SYSTEM

For data leakage detection and prevention, we will be comparing the optimal K-anonymity and incognito algorithm and compare both algorithms on bases of Minimal Distortion, Information Loss Caused After Data is Anonymized, Calculation of execution time and record the results of the algorithms performance in different cases of data leakage. Different tests on algorithms provides results that can be favourable for different purpose. This project will mostly focus on performance and outcome comparison of the results provided by the two algorithms.

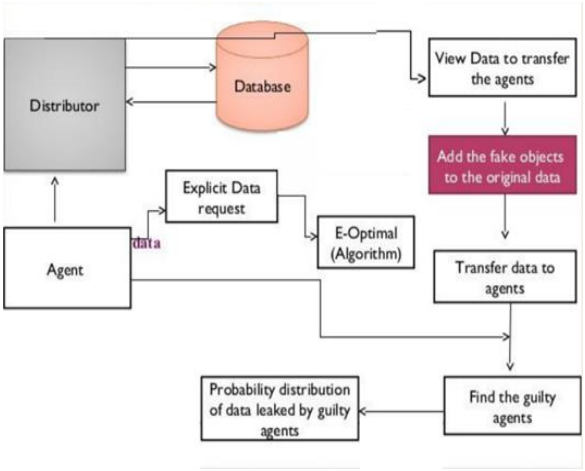

Fig 1. Block diagram of Proposed System

This is the general flow of a data leakage detection model. In this the distributor receive explicit data request from an agent, the distributor fetches the data from the databases and adds fake objects to the data request which the agent has requested. Since distributor can't modify the requested data by the agent. We use the E-Optimal algorithm to find fake object. The fake object also helps in identifying the guilty agent. Then the distributor forwards the original data with fake objects to agents. If agent leaks the data by any chance, here sample data request algorithm comes into play the main goal of this algorithm is to find the guilty agent, it takes probabilistic approach that is it uses s-overlap algorithm which minimizes sum objective which lets us guess the agent who might have licked the data. this is what a data leakage detection is. On top of this we implement the k-anonymity algorithm which adds a layer of privacy to the data sent.

4. EXPERIMENTAL RESULTS

NO.	NAME	AGE	SEX	DONOR CMV	SSN	BLOOD TYPE	ZIP CODE
1	Anthony	20-60	*	A1254	548456	A+	8****
2	Brian	20-60	*	B2564	987654	O-	80117
3	Charles	20-60	*	C5468	894154	B-	702**
4	David	20-60	*	D3846	129468	AB+	7****
5	Edward	20-60	*	E2312	484654	B+	8****
6	Frank	20-60	*	F7854	168458	AB-	8016*
7	Alice	20-60	*	G6520	895548	AB+	703**
8	Barbara	20-60	*	H1357	894815	O+	8016*
9	Carol	20-60	*	J5654	785456	A+	80117
10	Donna	20-60	*	K4698	484568	B-	703**
11	Emily	20-60	*	L5456	556454	B+	8016*
12	Fiona	20-60	*	M8474	784655	AB-	702**
13	Lawrance	20-60	*	G6524	236655	AB-	80182
14	Khalid	20-60	*	E2316	784865	O+	80182
15	Trivian	20-60	*	F7866	584675	AB+	80182
16	Nikki	20-60	*	W9989	974999	A-	60182
17	Abigale	20-60	*	U7545	153246	O+	80182
18	Elsa	20-60	*	Q6321	484655	B-	60182
19	Ava	20-60	*	L0012	853354	A-	80182
20	Eric	20-60	*	P5001	956651	O-	7****

Fig 2: File with Anonymized data after applying K-anonymity algorithm for k = 2.

In fig 2 the data in the file that is fed into the K anonymity algorithm will anonymize the data for k = 2. The value of k can be adjusted if required.

NO.	NAME	AGE	SEX	DONOR CMV	SSN	BLOOD TYPE	ZIP CODE
1	Anthony	3*	M	A1254	548456	A+	8****
2	Brian	2*	M	B2564	987654	O-	8****
3	Charles	4*	M	C5468	894154	B-	7****
4	David	3*	M	D3846	129468	AB+	7****
5	Edward	4*	M	E2312	484654	B+	8****
6	Frank	4*	M	F7854	168458	AB-	8****
7	Alice	3*	F	G6520	895548	AB+	7****
8	Barbara	2*	F	H1357	894815	O+	8****
9	Carol	3*	F	J5654	785456	A+	8****
10	Donna	5*	F	K4698	484568	B-	7****
11	Emily	4*	F	L5456	556454	B+	8****
12	Fiona	3*	F	M8474	784655	AB-	7****
13	Lawrance	5*	M	G6524	236655	AB-	8****
14	Khalid	4*	M	E2316	784865	O+	8****
15	Trivian	2*	M	F7866	584675	AB+	8****
16	Nikki	3*	F	W9989	974999	A-	6****
17	Abigale	3*	F	U7545	153246	O+	8****
18	Elsa	3*	F	Q6321	484655	B-	6****
19	Ava	3*	F	L0012	853354	A-	8****
20	Eric	5*	M	P5001	956651	O-	7****

Fig 3: File with anonymized data after applying Incognito algorithm for k = 2 (Most optimum preference).

In fig 3 the data in the file that is fed into the incognito algorithm to produce most optimum anonymized data will anonymize the data for k = 2. The value of k can be adjusted if required. If the distributor wishes to choose most secure anonymized data formation preference the data that will be anonymized will not provide agents with excessive information gain from the anonymized data but also simultaneously provide highest possible security to prevent data leakage for the distributor. The anonymized data after selecting the most secure preference the data will look similar to data if file from fig 4.

NO.	NAME	AGE	SEX	DONOR CMV	SSN	BLOOD TYPE	ZIP CODE
1	Anthony	**	*	A1254	548456	A+	*****
2	Brian	**	*	B2564	987654	O-	*****
3	Charles	**	*	C5468	894154	B-	*****
4	David	**	*	D3846	129468	AB+	*****
5	Edward	**	*	E2312	484654	B+	*****
6	Frank	**	*	F7854	168458	AB-	*****
7	Alice	**	*	G6520	895548	AB+	*****
8	Barbara	**	*	H1357	894815	O+	*****
9	Carol	**	*	J5654	785456	A+	*****
10	Donna	**	*	K4698	484568	B-	*****
11	Emily	**	*	L5456	556454	B+	*****
12	Fiona	**	*	M8474	784655	AB-	*****
13	Lawrance	**	*	G6524	236655	AB-	*****
14	Khalid	**	*	E2316	784865	O+	*****
15	Trivian	**	*	F7866	584675	AB+	*****
16	Nikki	**	*	W9989	974999	A-	*****
17	Abigale	**	*	U7545	153246	O+	*****
18	Elsa	**	*	Q6321	484655	B-	*****
19	Ava	**	*	L0012	853354	A-	*****
20	Eric	**	*	P5001	956651	O-	*****

Fig 4: File with anonymized data after applying Incognito algorithm for k = 2 (Most secure preference)

In fig 3 the data in the file that is fed into the incognito algorithm to produce most optimum anonymized data will anonymize the data for k = 2. The value of k can be adjusted if required.

The advantage that incognito algorithm provides helps distributor select what should be the quality of data anonymized. Whether they should be highly optimal so that the agents who receive this data can gain as much information as possible or if only a limited amount of information is to be gained by agents from this data but provide the highest level of data security anonymization can create for the distributor to prevent data leakage.

A. *Information Loss*

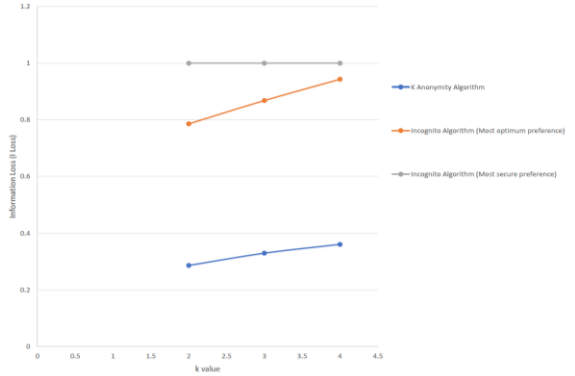

Fig 5: Information Loss result comparison graph

In fig 5 it is evident that with increase in the value of k information loss steadily increases when K anonymity algorithm and incognito algorithm with selection of most optimum preference selection is applied. When incognito algorithm with selection of most secure preference is selected, the information loss remains constant since most secure anonymized data does not provide any viable data for information gain which implies that there is complete loss of data.

B. *Minimal Distortion*

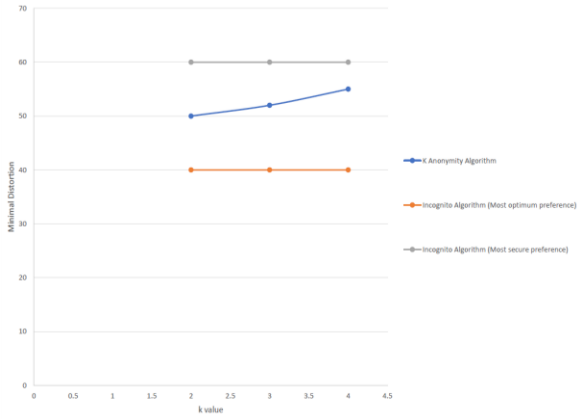

Fig 6: Minimal distortion result comparison graph

In fig 6 it is evident that with increase in the value of k minimal distortion steadily increases when K anonymity algorithm is applied and remains constant when incognito algorithm with selection of most optimum preference selection or with selection of most secure preference is applied. Incognito algorithm with any of the two preferences may have constant minimal distortion but the distortion in data is higher for anonymized data when most secure preference is selected. Anonymized data when incognito algorithm with selection of most optimum preference is selected it will have lower distortion in data when compared to data anonymized with K anonymity.

C. *Execution Time*

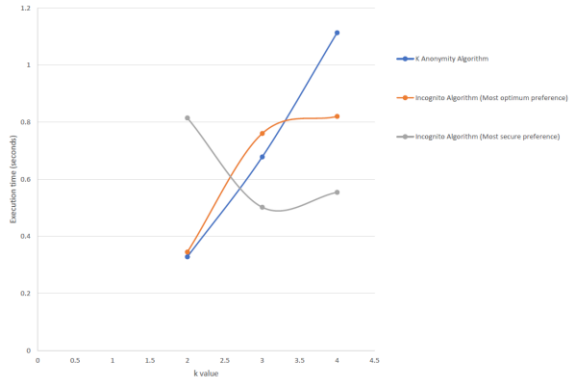

Fig 7: Execution Time result comparison graph

In fig 7 it is evident that execution time to anonymize data using K anonymity algorithm and incognito algorithm with selection of most optimum preference increases with increase in k value. But in case of

incognito algorithm with selection of most secure preference the execution time decreases to a point and then there is a slight increase since most secure form of anonymized data using incognito algorithm does not have change drastically and remains the same even with increase in k value.

D. Result

Information loss caused after data is anonymized.	Incognito algorithm (Most secure preference)	> Incognito algorithm (most optimum preference)	> K anonymity algorithm
Minimal Distortion of data anonymized.	Incognito algorithm (Most secure preference)	> K anonymity algorithm	> Incognito algorithm (most optimum preference)
Execution time of the algorithms.	K anonymity algorithm	> Incognito algorithm (most optimum preference)	> Incognito algorithm (Most secure preference)

Fig 8: Result Summarization table

5. CONCLUSION

Here the main focus of our paper is the data allocation problem, how can the distributor "intelligently" give data to agents in order to improve the chances of detecting a guilty agent. We have formally defined the problem of k-anonymizing a database via suppressing tuple components, and determined the computational complexity of k-anonymization when one wishes to withhold a minimum number of entries yet achieve a privacy level k. For data leakage detection we use the sample data and explicit data request to determine the guilty agent using the fake objects. For preventing data leakage detection, we utilize either k anonymity or the incognito algorithm to generalize and suppress the data. When we utilize these prevention algorithms, we need to take into consideration the information loss of data whether its required to be high or low and same conditions are used for data distortion. After mathematically evaluating the generalized and suppressed data we find that information loss is greater in data when incognito algorithm is applied to acquire anonymized data with most security and data distortion is minimal in k anonymity algorithm but substantially greater when compared with incognito algorithm if the most optimum data anonymization preference is chosen

REFERENCES

[1] Data Leakage Detection By Using Fake Objects By Rama Rajeswari Mulukutla & P. Poturaju Grandhi Varalakshmi Venkatarao Institute Of Technology, India,Volume 13 Issue 6 Version 1.0 Year 2013 Publisher: Global Journals Inc. (Usa).

[2] A Systematic Comparison and Evaluation Of K-Anonymization Algorithms for Practitioners Vanessa Ayala-Rivera∗, Patrick Mcdonagh, Thomas Cerqueus, Liam Murphy, Transactions on Data Privacy 7 (2014) 337–370.

[3] Incognito: Efficient Full-Domain K-Anonymity, Kristen Lefevre, David J. Dewitt, Raghu Ramakrishnan,Sigmod '05: Proceedings of the 2005 Acm Sigmod International Conference on Management of Data June 2005 Pages 49–60.

[4] (A, K)-Anonymity: An Enhanced K-Anonymity Model for Privacy Preserving Data Publishing, Raymond Chi-Wing Wong, Jiuyong Li, Ada Wai-Chee Fu, Ke Wang,Kdd '06: Proceedings of the 12th Acm Sigkdd International Conference On Knowledge Discovery And Data Miningaugust 2006 Pages 754–759.

[5] "Data Leakage Detection Using K-Anonymity Algorithm", Ms.B.Kohila, Mrs.K.Sashi, International Journal Of Computer Science And Management Research Vol 1 Issue 5 December 2012

[6] Emam Et Al., A Globally Optimal K-Anonymity Method for The De-Identification of Health Data, Khaled El Emam, Fida Kamal Dankar, Romeo Issa, Elizabeth Jonker, Daniel Amyot, Elise Cogo, Jean-Pierre Corriveau, Mark Walker, Sadrul Chowdhury, Regis Vaillancourt, Tyson Roffey, Jim Bottomley, Journal of The American Medical Informatics Association Volume 16 Number 5 September / October 2009.

[7] Preserving Privacy During Big Data Publishing Using K-Anonymity Model – A Survey, Divya Sadhwani, Dr.Sanjay Silakari, Mr. Uday Chourasia, International Journal Of Advanced Research In Computer Science, Volume 8, No. 5, May-June 2017.

[8] "K-Anonymity, V. Ciriani, S. De Capitani Di Vimercati, S. Foresti, And P. Samarati,Secure

Data Management in Decentralized Systems (Pp.323-353),January 2007

[9] "Data Anonymization - Generalization Algorithms", Li Xiong, Slawek Goryczka Cs 573 Data Privacy and Security - Hippocratic Database Technology, Spring 2012.

[10] Research on Behavior-Based Data Leakage Incidents for The Sustainable Growth Of An Organization Jawon Kim 1, Jaesoo Kim 2 And Hangbae Chang 3. Computing, August.