International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 04 Issue-17
December 2017

Data Linkage and Leakage Detection in Data Mining Using E-Random and S-Random

Mr. SreedharAmbala , Mr.Mamidala Sagar , Ms.K.Lavanya
[1,2]Asst.Professor. Siddhartha Institute of Technology and Sciences, Hyderabad.
[3]Pursuing M.Tech Siddhartha Institute of Technology and Sciences, Hyderabad.
Email: sridhar.Ambala@gmail.com , Email: Mamidala.sagar@gmail.com , Email: lavanya701@gmail.com

Abstract:

A data distributor has given sensitive data to a set of supposedly trusted agents (third parties). Some of the data has leaked and found in an unauthorized place (e.g., o n the web or somebody's laptop). The distributor should assess the likelihood of the leaked data came from one or more agents, as opposed to having independently gathered by others. We propose data allocation strategies (across the agents) that improve the probability of identifying leakages and linkage the data. These methods don't rely on alterations of the released data (e.g., watermarks). In some cases we can also inject "realistic but fake" data records to further improve our chances of detecting leakage and identifying the guilty party. The main objective of this paper E-Random algorithm and S-Random algorithm is used by adding additional information (i.e. Fake data) to the original data to detect the abnormal access in database records very effectively.

Keywords: E -Random, S-Random, M -score, guilty agent, Distributer.

I. INTRODUCTION:

Ever since in Data Linkage techniques, there is a dramatic challenge for identifying the fraud detection information illegally over the internet (i.e. license key, Applications, Ban k Transaction Details etc..,) to different unauthorized users. It makes critical problem for handling confidential and many other secret information. In existing system, many techniques can be used to for detecting illegal distributing users. In [4] One-Class Clustering Tree (OCCT) method fraud detection is used to detect the abnormal access to database records that might indicate a data misuse. The main disadvantage of this method is that it takes more time to detect data misusage. In proposed system, E-Random algorith m and S-Rando m algorith m is used by adding additional informat ion (i.e. Fake data) to the original data. This method is used to detect the abnormal access in database records very effectively and it takes less time to detect the illegal file distribution compared to OCCT.

Our aim is to detect when the distributor's sensitive data have been leaked by agents, and if possible to identify the particular agent that leaked the data. Perturbation is most useful technique where the data has modified and made "less sensitive" before being watermarking is used to handle the leakage detection. We annunciate the need for watermarking database relations to deter their piracy, and identify the unique characteristics of relational data which pose new challenges for watermarking, and provide desirable properties of watermarking system for relational data. A watermark can be applied to any of the database relation having attributes which are such that changes in a few of their values do not affect the applications. Watermarking means a unique code is embedded in each distributed copy If that copy is later discovered

in the hands of an unauthorized party, the leaker can be identified. Furthermore, watermarks can sometimes be destroyed if the data recipient is malicious.

2. LITERATUR E S URVEY:

The Paper [1] used to represented a one class clustering tree approach which performs one-to-many record linkage. This method is based on a one class decision tree model which sums up the knowledge of which records to be linked together.

To summarize, this method allows performing one-to-many linkage while the traditional methods followed one-to-one linkage. Then, we have used a one-class approach which results in matching pairs are only required in the training set, as more nu mber of non-matching (negative) pairs will confuse the model and it will lead to a less accurate model. Another advantage of using OCCT model is that the solution can be easily transformed to ru les.

The paper [2] proposes the vast amount of informat ion on the World Wide Web, a typical short query of 1-3 words submitted to a search engine usually get a result list of tens of thousands web pages, while only a t iny part of these pages is useful for users. Web data cleansing with key resource selection based on K-means clustering makes it possible to get better retrieval performance with fewer pages indexed.

This paper [3] has been to demonstrate that the technology for building call trees fro m examp les is fairly sturdy. Current industrial systems square measure powerful tools that have achieved noteworthy successes. The groundwork has been fin ished advances that may allow such tools to deal even with shouting, incomplete informat ion typical of advanced real-world applications. Work is constant at many centers to enhance the performance of the underlying algorith ms. 2 samples of modern analysis provide some tips to the directions during which the sector is moving. Where as call trees generated by the higher than systems square measure quick to execute and might be terribly correct, they leave a lot of to be desired as representations of knowledge.

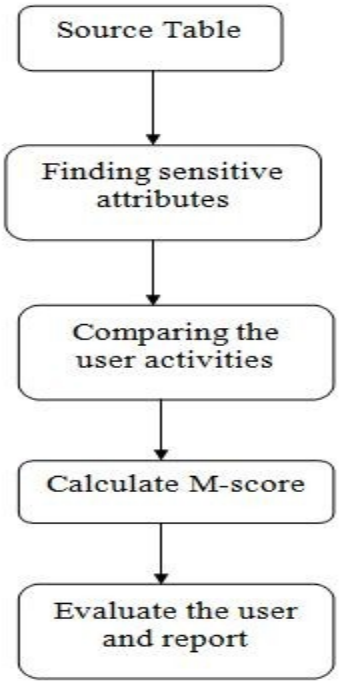
3. SYS TEM ARCHITECTURE:

Fig 1 and Fig 1.1 shows overall flow of our proposed system. A knowledge distributor has given sensitive data to a collection of purportedly sure agents (third parties). A number of the in formation is leaked and located in AN unauthorized p lace (e.g., on

the net or somebody's laptop). The distributor should assess the chance that the leaked knowledge came fro m one or additional agents, as critical having been severally gathered by different suggests that. We have a tendency to propose knowledge allocation ways (across the agents) that improve the chance of distinguishing leakages. These ways don't have confidence alterations of the discharged knowledge (e.g., watermarks). In so me cases we are able to additionally inject "realistic however

fake" knowledge records to additional improve our probabilit ies of police work discharge and distinguishing the guilty one. Our goal is to notice once the distributor's sensitive knowledge has been leaked by

Fig 1.1 system flow diagramagents, and if doable to spot the agent that leaked the informat ion.



The main focus of our project is that the knowledge allocation drawback as however will the distributor "intelligently" provides knowledge to agents so as to enhance the probabilities of police work a guilty agent.

Fake objects are objects generated by the distributor so as to extend the probabilities of police work agents that leak knowledge.

The distributor could also be able to add faux objects to the distributed knowledge so as to enhance his effectiveness in police work guilty agents. Our use of fau x objects is impressed by the employment "trace"

4. EXISTING S YSTEM:

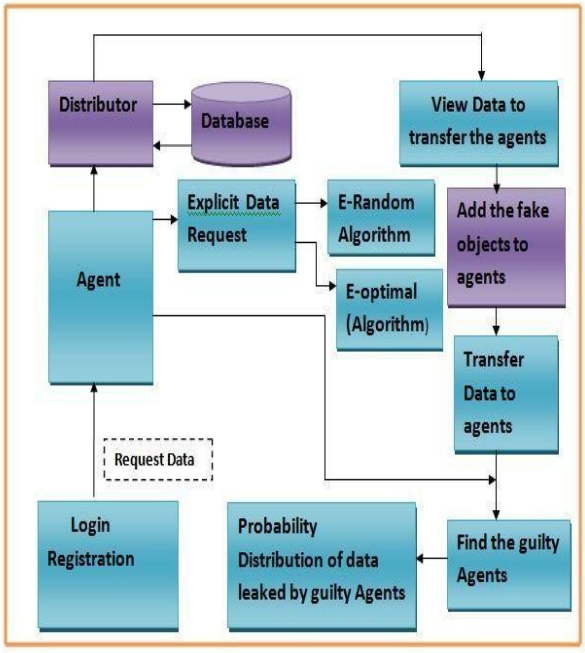
- One-Class Clustering Tree (OCCT) method[4] is used for fraud detection in data lin kage.
- Datasets are constructed in tree structures.
- Datasets is splitted into number of subsets to perform co mparison using CGJ (Coarse-Grained Jaccard).

4.1 DISADVANTAGES

- Tree is constructed in small amount of nodes.
- It is complex to process large number of nodes.
- It takes more t ime to co mpare the data lin kage in large nu mber o f subsets.

5. PROPOS ED S YSTEM:

- In proposed system, we present a new concept, Misuseability Weight, for estimating the risk emanating fro m data exposed to insiders.
- This concept focuses on assigning a score that represents the sensitivity level of the



records in mailing lists.

data exposed to the user and by that predicts the ability of the user to maliciously exploit this data.

- E-Rando m algorith m and S-Random algorith m is used by adding additional

  informat ion (i.e . Fake data) to the original data and randomly allocating fake objects to the agent.

5.1 ADVANTAGES

- It is more flexible to process large number

  of nodes.

- It takes less time to compare the data lin kage in large number of subsets.

- It is easy to detect and calculate the abnormal access in the database.

6. SYS TEM IMPLEMENTATION:

6.1. DATA ALLOCATION PROBLEM

A. Fake objects:

Fake objects are objects generated by the distributor in order to increase the chances of detecting agents that leak data. The d istributor may be add fake objects to the distributed data in order to improve his effectiveness in detecting guilty agents. Our use of fake objects is inspired by the use of

"trace" records in mailing lists . The idea of perturbing data to detect leakage is not new,e.g., [1]. However, in most cases, individual objects are Perturbed, e.g., by adding random noise to sensitive salaries,or adding fake elements .

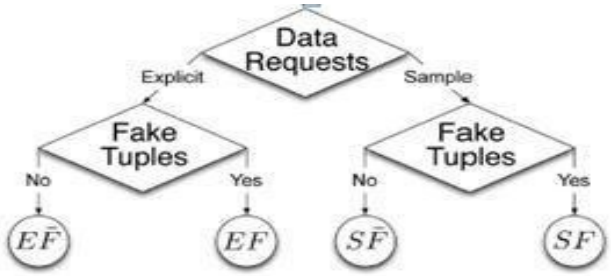


Fig 2. Leakage problem Instances

In some applications, fake objects may cause fewer problems that perturbing real objects. Creation. The creation of fake but real-looking objects is a nontrivial problem whose thorough investigation is beyond the scope of this paper. Here, we model the creation of a fake object for agent Ui as a black bo x function CREATEFAKEOBJECT (Ri, Fi, condi) that takes as input the set of all objects Ri, the subset of fake objects Fi that Ui has received so far, and condi, and returns a new fake object.This function needs condi to produce a valid object that

satisfies Ui's condition. Set Ri is needed as input so that the created fake object is not only valid but also indistinguishable from other real objects .

Although we do not deal with the implementation of CREATEFA KEOBJECT (), we note that there are two main design options. The function can either produce a fake object on demand every t ime it is called or it can return an appropriate object from a pool of objects created in advance.We are using the following strategies to add the fake object to finding guilty agent.

B. optimization problem

The Optimization Module is the distributor's data allocation to agents has one constraint and one objective. The distributor's constraint is to satisfy agents' requests, by providing them with the number of objects they request or with all available objects that satisfy their conditions. His objective is to be able to detect an agent who leaks any portion of his data.We consider the constraint as strict. The distributor may not deny serving an agent request and may not provide agents with different perturbed versions of the same objects as in [1]. We consider fake object d istribution as the only possible constrain

relaxation. Our detection objective is ideal and intractable.

A data leakage is the intentional or unintentional release of secure information to an untrusted environment. Other terms for this phenomenon include unintentional informat ion disclosure, databreach and also data spill. Incidents range from concerted attack by black hats with the backing of organized crime or national governments to careless disposal of used computer equipment or data storage media.

C. Objective Approximation

We can approximate the objective of (2) with (3) that does not depend on agents' guilt probabilit ies, and therefore, on **p**:

Maximise (...|Ri∩Rj|/|Ri|,....) i ≠j. .....(3)

This approximat ion is valid if min imizing the relative overlap

|Ri∩Rj|/|Ri| maximizes Δ (i ,j).

Therefo re, we can scalarize the problem objective by assigning the same weights to all vector objectives.

Maximize ∑i=1 1/|Ri|∑j=1|Ri ∩ Rj| ..... (4a)
(Over R1... Rn)

Maximize max |Ri ∩ Rj| /|Ri| ..... (4b)
(Over R1... Rn)

Both scalar optimization problems yield the optimal solution of the problem of (3), if

such solution exists. If there is no global optimal solution, the sum-objective y ields

the Pareto optimal solution that allows the distributor solution that guarantees that the distributor will detect the guilty agent with certain confidence in the worst case. Such guarantee may adversely impact the average performance of the distribution.

6.2. ALLOCATION STRATEGIES

The main focus of our project is the data allocation problem as how can the distributor "intelligently" give data to agents in order to imp rove the chances of detecting a guilty agent.

6.3.ALGORITHM US ED

6.3.1 Explicit and Sample Random Algorithm

A.Explicit Data Requests

In problems of class EF, the distributor is not

allo wed to add fake objects to the distributed data.

So, the data allocation is fully defined by the agents' data requests. Therefore, there is nothing to optimize. In EF problems, The d istributor cannot remove or alter the R1 or R2 data to decrease the overlap R1∩R2. However, say that the distributor can create one fake object (B= 1) and both agents can receive

one fake object (b1= b2= 1). In this case, the distributor can add one fake object to either R1 or R2 to increase the corresponding denominator of the summation term. Assume that the distributor creates a fake object f and he gives it to agent R1. Agent U1 has now R1= {t1, t2, f}and F1={f} and the value of the sum-object ive decreases to 1/3+1/1=1.33<1.5.

Algorithm 1. Allocation for Explicit Data Requests (EF)
Input: R1. . . Rn, cond1; . . . ; condn, b1,. . . , bn, B
Output: R1. . . Rn, F1. . . Fn
1: R ←ϕ Agents that can receive fake objects
2: for i =1... n do

3: if bi > 0 then
4: R←R∪ {i}
5: Fi ←ϕ ;
6: while B > 0 do
7: i ← SELECTAGENT(R, R1. . .Rn)
8: f ← CREATEFAKEOBJECT (Ri, Fi, condi)
9: Ri ← Ri ∪ {f}
10: Fi ← Fi ∪{f}
11: bi ← bi _ 1
12: if bi = 0then
13: R← R/ {Ri}
14: B← B _ 1

Algorithm 2. Agent Selection for e-random
1: function SELECTAGENT (R, R1; . . ., Rn)
2: i← select at random an agent from R
3: return i

In lines 1-5, Algorithm 1 finds agents that are elig ible to receiving fake objects in O(n) time. Then, in the main loop in lines 6-14, the algorithm creates one fake object in every iteration and allocates it to random agent. The main loop takes O (B) t ime. Hence, the running time of the algorith m is

O(n+ B). If B $\geq \Sigma n$ i=1 b i, the algorith m minimizes every term of the objective summat ion by adding the maximu m number bi of fake objects to every set Ri, yielding the optimal solution. Otherwise, if B < Ei=1 bi (as in our examp le where B=1 < b1+b2 =2), the algorith m just selects at random the agents that are provided with fake objects. We return back to our example and see how the objective would change if the distributor adds fake object f to R2 instead of R1. In this case, the sum-objective would be 1/2+1/ 2=1<1.33. The reason why we got a greater improvement is that the addition of a fake object to R2 has greater impact on the corresponding summation terms, since

1/|R1|-1/ |R1|+1=1/ 6<1/|R2|-1/|R2|+1=1/2.

The left-hand side of the inequality corresponds to the objective improvement after the addit ion of a fake object to R1 and the right-hand side to R2.

Algorith m 3. Agent Selection for e-optimal
1: function SELECTA GENT (R,R1; . . .; Rn)
2: i← argmax $(1/|Ri'|-1/|Ri'|+1)\Sigma|Ri'\cap Rj|$
3: return i

Algorith m 3 makes a greedy choice by selecting the agent that will yield the greatest improvement in the sum objective

the cost of this greedy choice is O (n2) in every iteration.The overall running time of e-optimal is O (n+n2B) =O(n2B). Theorem 2 shows that this greedy approach finds and optimal d istribution with respect to both optimization objectives defined in (4).

Theorem 2. Algorith m e-optimal yields an object allocation that min imizes both sum and max-objective in problem instances of class EF.

B.Sample Data Requests:

With sample data requests, each agent Ui may receive anyT subset out of (|T|) d ifferent ones. Hence, there are i=1(|T|) different object allocations. In every allocation, the distributor can permute T objects and keep the same chances of guilty agent detection. The reason is that the guilt

probability depends only on which agents have received the leaked objects and not on the identity of the leaked objects.Therefore, fro m the distributor's perspective, different

allocations. The distributor's problem is to

pick one out so that he optimizes his objective. We formulate the problem as a nonconvex QIP that is NP-hard.

Note that the distributor can increase the number of possible allocations by adding fake objects (and increasing |T|) but the

problem is essentially the same. So, in the rest of this section, we will on ly deal with problems of class SF, but our algorithms are applicab le to SF problems as well.

C. Random: An object allocation that satisfies requests and ignores the distributor's objective is to

We present S-random in two parts: Algorithm 4 is a general allocation algorith m that is used by other algorith ms in this section. In line 6 of A lgorith m 4, there is a call to function SELECTOBJECT () whose implementation differentiates algorithms that rely on Algorith m 4. Algorithm5 shows function SELECTOBJECT () for s -rando m. A lgorith m 4. Allocation for Samp le Data Requests (SF)

Input: m1, . . .,mn, |T| . Assuming mi <|T|
Output: R1, . . .,Rn
1: a← 0|T|. a[k]:number of agents who have received object tk
2: R1 ←ϕ. . . Rn ←ϕ
3: remaining ←Σi=1 mi
4: while remaining > 0 do
5: for all i = 1. . . n: |Ri| <mi do
6: k← SELECTOBJECT (i, Ri) May also use additional parameters
7: Ri ← Ri ∪{tk}
8: a[k] ← a[k] + 1
9: remaining ←remaining _ 1
Algorithm 5. Object Selection for s-random
1: function SELECTOBJECT (i, Ri)
2: k ←select at random an element from set {K'/tk∉Ri}
3: return k

7.CONCLUS ION:

In this paper We have shown that it's attainable to assess the likelihood that associate agent is chargeable for a leak, based on the overlap of his knowledge with the leaked knowledge and therefore the data of different agents, and supported the likelihood that objects are often "guessed" by give each agent Ui a rando mly selected subset of T of size mi. We denote this algorithm by S-random and we use it as our baseline. different suggests that. We are proposing our increased approach for detection the guilty agents. during this technique we have a tendency to use the linguistics inference model that represents the likelihood of possible colluding attacks from any agents to the different knowledge allocation methods. we present a new concept, Misuseability Weight, for estimating the risk emanating fro m data exposed to insiders. In this we are calcu lating Misuseability weight and probability value based on how much data misused or leakage data by agents. so that distributor can easily estimate misused data and guilty agent.

In future the extension of our allocation methods will handle agent requests in an internet fashion (the presented methods assume that there's a hard and fast set of agents with requests proverbial in advance) are often implemented.

REFERENCES

[1] Sunandhini.S and Suguna.M," Improved One-to-Many Record Lin kage using One Class Clustering Tree", (IJCA)(0975– 8887) ,International Conference on Simu lations in Co mputing Nexus, ICSCN-2014.
[2] YiqunLIU,MinZHANG, Canhui WANG, Shaoping MA,"Learning-based Web Data Cleansing for Information Retrieval" , Journal of Co mputational Information Systems 1:2 (2005) 203-213,june 200
[3] J.R. QUINLAN,"Induction of Decision Trees", Machine Learn ing 1: 81-106, 1986
[4] Ma'ayan Dror,Asaf Shabtai,Lior Rokach and Yuval Elovici, "OOCT:A One class clustering Tree for implementing One-to- many Data lin kage" IEEE,vol.26,ZNo.3,March 2014.
[5] B.Mungamuru and H.Garcia mo lina,"privacy,preservation and Performance: The 3 p's of Distributed Data Management," technical report, Stanford univ.,2008
[6] R. Agrawal and J. Kiernan, "Watermarking Relational Databases,"Proc. 28th Int'l Conf. Very Large Data Bases (VLDB '02), VLDB Endowment, pp. 155-166, 2002.
[7] L. Sweeney, "Achieving K-Anonymity Privacy Protection Using Generalization and Suppression," http://en.scientificco mmons. org/43196131, 2002.