Fake objects allocation for data leakage detection

Odela Sandeep Kumar[1], B.Sunitha Devi [2]

[1]Student, Dept of CSE, CMR Institute of Technology,Hyderabad,AP,INDIA

[2]Associate ProfessorDept of CSE,CMR Institute of Technology,Hyderabad,INDIA,Hyderabad,AP,INDIA

Abstract:

An information distributor has given sensitive data to a collection of purportedly trust worthy agents (third parties). A number of the information has leaked and located in an unauthorized place (e.g., on the net or somebody's laptop). The distributor ought to assess the probability of the leaked information came from one or a lot of agents, as opposition having severally gathered by others. We tend to propose information allocation methods (across the agents) that improve the likelihood of characteristic leakages. These ways don't consider alterations of the discharged information (e.g., watermarks). In some cases we are able to additionally inject "realistic however fake" information records to any improve our probabilities of detective work run and characteristic the problem.

Key Words: Data leak, information misuse, security measures, misuse ability weight.

1. INTRODUCTION:-

In the course of doing a business, typically sensitive information should be bimanual over to purportedly trustworthy agents. As an example, a corporation could have partnerships with the opposite corporations hat need sharing client information. Another enterprise may source its processing, so the information should run to numerous alternative corporations. Our aim is to notice once the distributor's sensitive information is leaked by agents, and if potential to spot the actual agent that leaked the information. Perturbation is most helpful technique wherever the data has changed and created "less sensitive" before being bimanual to agents For example, one will replace the precise. Values by ranges or one will add the random noise to sure attributes. Historically, watermarking is employed to handle the outpouring detection. We have a tendency to foretell the necessity for watermarking information relations to discourage their piracy, and establish the distinctive characteristics of relative information that create new challenges for watermarking, and supply fascinating properties of watermarking system for relative information. A watermark may be applied to any of the information relation having attributes that ar specified changes in a very few of their values don't have an effect on the applications. Watermarking suggests that a novel code is embedded in every distributed copy If that duplicate is later discovered within the hands of associate unauthorized party, the informant may be known. Moreover, watermarks will typically be destroyed if the information recipient is malicious

In this paper, we have a tendency to study techniques for detective work outpouring of a collection of objects or records. Specifically, we have a tendency to study the subsequent scenario: when giving a collection of objects to agents, the distributor discovers a number of those self same objects in associate unauthorized place. At now, the distributor will assess the probability that the leaked information came from one or a lot of agents, as against having been severally gathered by alternative suggests that data, he could stop doing business with him, or could initiate legal proceedings.

In this paper, we have a tendency to develop a model for assessing the "guilt "of agents. We have a tendency to conjointly gift algorithms for distributing objects to agents, in a very means that improves our possibilities of characteristic a informant. Finally, we have a tendency to conjointly think about the choice of adding "fake" objects to the distributed set. Such objects don't correspond to real entities however seem realistic to the agents. In a sense, the faux objects act as a sort of watermark for the complete set, while not modifying any person members.

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

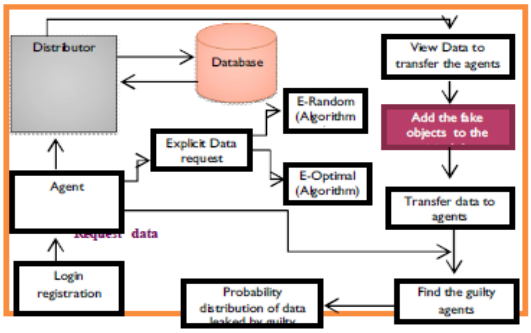


Fig1.System Architecture

II. DATA ALLOCATION PROBLEM

A. Fake objects:

Fake objects area unit objects generated by the distributor so as to extend the possibilities of detective work agents that leak information. The distributor could also be adding pretend objects to the distributed information so as to boost his effectiveness in detective work guilty agents. Our use of pretend objects is galvanized by the utilization of "trace" records in mailing lists. the thought of disturbing information to observe escape isn't new, e.g., [1]. However, in most cases, individual objects area unit hot and bothered, e.g., by adding random noise to sensitive salaries, or adding pretend parts.

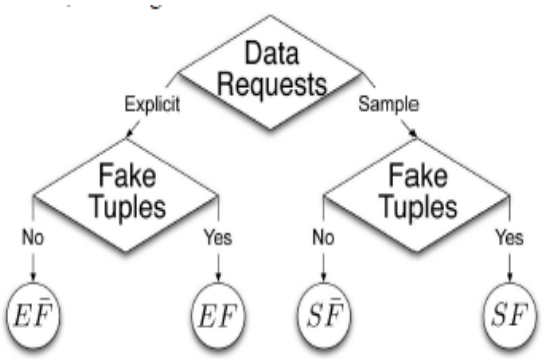


Fig2. Leakage problem instances.

Fig-3: Distributor Page

In some applications, pretend objects might cause fewer issues that disturbing real objects Creation. The creation of pretend however real-looking objects could be a nontrivial downside whose thorough investigation is on the far side the scope of this paper. Here, we have a tendency to model the creation of a pretend object for agent Ui as a recorder operate CREATEFAKEOBJECT (Ri, Fi, condi) that takes as input the set of all objects RI, the set of pretend objects Fi that Ui has received to date, and condi, and returns a brand new pretend object.This operate desires condi to provide a legitimate object that Satisfies Ui's condition. Set RI is required as input in order that the created pretend object isn't solely valid however conjointly indistinguishable from different real objects.

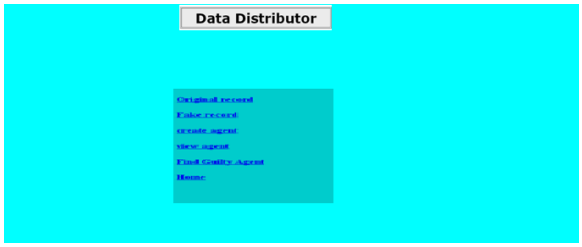


Fig-4: Distributor details

Though we have a tendency to don't touch upon the implementation of CREATEFAKEOBJECT (), we have a tendency to note that there are a unit 2 main style choices. The operate will either turn out a pretend object on demand whenever it's referred to as or it will come back associate applicable object from a pool of objects created ahead. We area unit mistreatment the subsequent methods to feature the pretend object to finding guilty agent

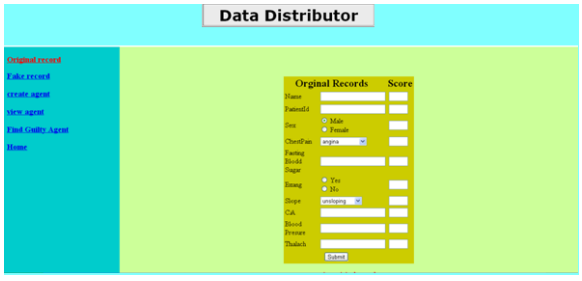


Fig-5: Adding the Original Records

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

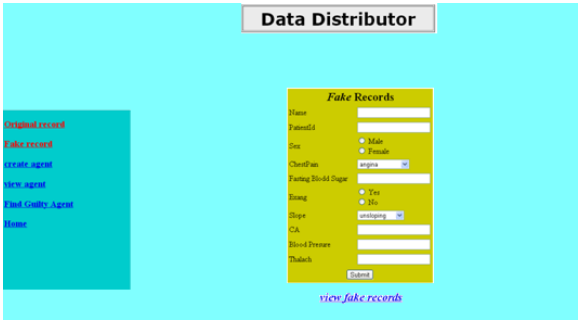


Fig4. Adding the Fake objects

B. optimization problem:

The optimisation Module is that the distributor's knowledge allocation to agents has one constraint and one objective. The distributor's constraint is to satisfy agents' requests, by providing them with the quantity of objects they request or with all on the market objects that satisfy their conditions. His objective is to be able to observe associate agent World Health Organization leaks any portion of his knowledge. We take into account the constraint as strict. The distributor might not deny serving associate agent request and should not offer agents with completely different hot and bothered versions of identical objects as in [1]. we tend to take into account faux object distribution because the solely doable constraint relaxation. Our detection objective is good and uncontrollable.

Detection would be assured providing the distributor gave no knowledge object to any agent. We tend to use instead the subsequent objective: maximize the possibilities of detection a guilty agent that leaks all his knowledge objects. we tend to currently introduce some notation to state formally the distributor's objective. Recall that Pr Ocean State} or just Pr is the likelihood that agent Uj is guilty if the distributor discovers a leaked table S that contains all Ri objects. we tend to outline the distinction functions Δ(i, j) as

Δ(i, j)=Pr-Pr i, j=1,…..n. ……(1)

Problem Definition. Let the distributor have knowledge requests from n agents. The distributor desires to administer tables R1, ..., Rn to agents U1, . . . , Un, severally, so that . He Satisfies agents' requests, and he maximizes the guilt likelihood variations Δ(I,j) for all i ,j=1and i≠ j. forward that the American state sets satisfy the agents' requests, we are able to categorical

The problem as a multicriterion optimisation

Problem:

Maximize (….,Δ(i ,j),..) i≠ j. …….. (2),

If the optimisation downside has associate optimum answer, it means there exists associate allocation D*= (R1*,…, Rn*} such the other possible allocation yields D*=(R1*,…,Rn*} yields Δ(i,j)≥Δ*(i,j) for all i; j. this implies that allocation Tj* permits the distributor to tell apart any guilty agent with higher confidence than the other allocation, since it maximizes the likelihood Pr with regard to the other likelihood Pr with j ≠i. though there's no optimum allocation D*, a multicriterion downside has sociologist optimum allocations.

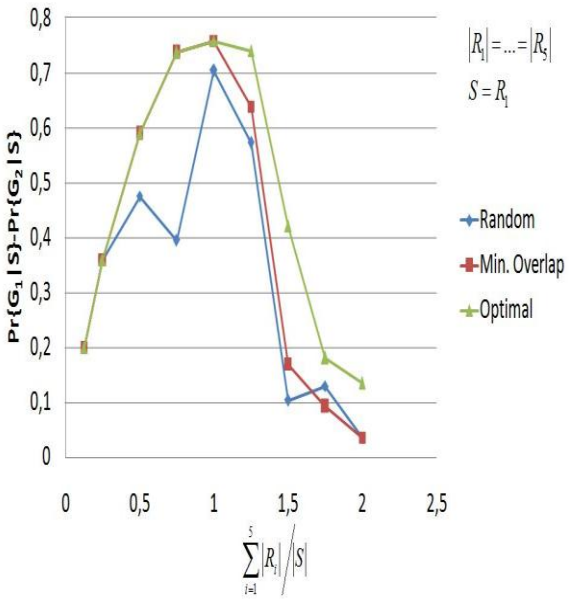
A data outflow is that the intentional or unintentional unharnessed of secure info to associate entrusted setting. Different terms for this development embrace unintentional info revealing, knowledge breach and additionally data spill. Incidents vary from conjunctive attack by black hats with the backing of social group or national governments to careless disposal of used laptop instrumentation or knowledge storage media.

C. Objective Approximation

We are able to approximate the target of (2) with (3) that doesn't rely on agents' guilt possibilities, and so, on p. This approximation is valid if minimizing the relative overlap Therefore; we are able to scalarize the matter objective by assignment identical weights to all or any vector objectives. Both scalar optimisation issues yield the optimum answer of the matter of (3), if such answer exists. If there's no world optimum answer, the sum-objective yields the sociologist optimum answer that enables the distributor to observe the guilty agent, on the average (over all completely different agents), with higher confidence than the other distribution. The max objective yields the answer that guarantees that the distributor can observe the guilty agent with bound confidence within the worst case. Such guarantee could adversely impact the common performance of the distribution.

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

EXPERIMENTAL RESULTS:



III. 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 improve the chances of detecting a guilty agent.

A. Explicit Data Requests

In problems of class EF, the distributor is not allowed 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, objective values are initialized by agents' data requests. Say, for example, that T= {t1, t2} and there are two agents with explicit data requests such\ that R1= {t1, t2} and R2 = {t1|.The value of the sum objective is in this case

$$\Sigma i=1\ 1/|Ri|\Sigma j=1|Ri \cap Rj|=1/2+1/1=1.5.$$

The distributor cannot remove or alter the R1 or R2 data to decrease the overlap $R1 \cap R2$. 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-objective 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 eligible 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) time. Hence, the running time of the algorithm is

O(n+ B). If B $\geq \Sigma n$ i=1 bi, the algorithm minimizes every term of the objective summation by adding the maximum number bi of fake objects to every set Ri, yielding the optimal solution. Otherwise, if B < Ei=1 bi (as in our example where B=1 < b1+b2 =2), the algorithm 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.

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

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

Algorithm 3. Agent Selection for e-optimal

1: function SELECTAGENT (R,R1; . . .;Rn)

2: i← argmax (1/|Ri′|-1/|Ri′|+1)Σ|Ri′∩Rj|
3: return i

Algorithm 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 an

Optimal distribution with respect to both optimization objectives defined in (4).

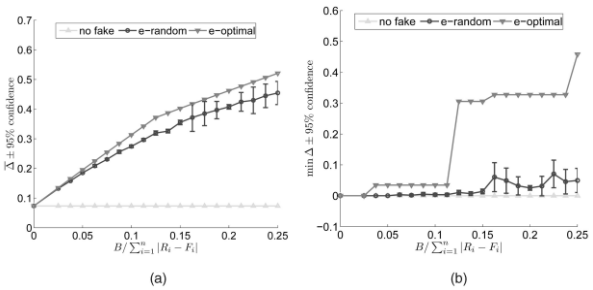
Theorem 2. Algorithm e-optimal yields an object allocation that minimizes both sum- and max-objective in problem instances of class EF.

B. Sample Data Requests

With sample knowledge requests, every agent Ui might receive any set out of (|T|) totally different ones. Hence, there are=1(|T|) totally different object allocations. In each allocation, the distributor will transpose T objects and keep similar probabilities of guilty agent detection. The rationale is that the guilt probability depends solely on that agents have received the leaked objects and not on the identity of the leaked objects. Therefore, from the distributor's perspective, different allocations. The distributor's downside is to pick one out so he optimizes his objective. We have a tendency to formulate the matter as a non convex QIP that's NP-hard.

Fig-6:Agent Page

Note that the distributor will increase the quantity of doable allocations by adding faux objects (and increasing |T|) however the matter is actually a similar. So, within the remainder of this section, we are going to solely agitate issues of sophistication SF, however our algorithms area unit applicable to SF issues further.



C. Random

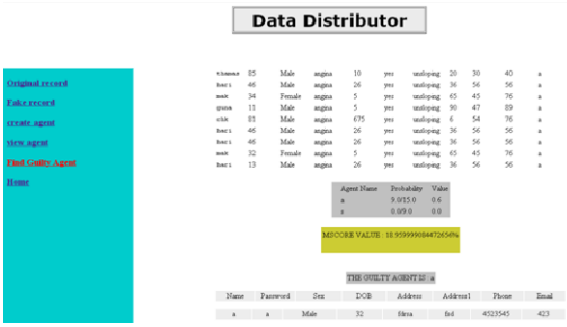
An object allocation that satisfies requests and ignores the distributor's objective is to convey every agent Ui a arbitrarily elect set of T of size mi. we tend to denote this rule by S-random and that we use it as our baseline. we tend to gift S-random in 2 parts: rule four could be a general allocation rule that's employed by different algorithms during this section. In line half-dozen of rule four, there's a decision to perform SELECT OBJECT() whose implementation differentiates algorithms that have faith in rule four. Rule five shows perform SELECTOBJECT () for s-random.

Algorithm for. Allocation for Sample information 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

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

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



ACKNOWLEDGMENT

We wish to acknowledge the efforts of **Pantech Solution Pvt ltd., Hyderabad,** for guidance which helped us work hard towards producing this research work.

REFERENCES

[1] A.subbiah and D.M.Blough.An Approach for fault tolerant and securedata storage in collaborative Work environments.
[2] .B.Mungamuru and H.Garcia molina,"privacy,preservation and Performance: The 3 p's of Distributed Data Management," technical report, Stanford univ.,2008
[3] M. Atallah and s.Wagstaff. Watermarking with quadratic residues. In
proc.of IS&T/SPIE Conference on Security and Watermarking of Multimedia Contents, January 1999.
[4] P. Buneman and W.-C. Tan, "Provenance in Databases," Proc. ACM SIGMOD, pp. 1171-1173, 2007
[5] S.Katzenbeisser and F.A.peticolas,editors. Information Hiding
Techniques for Steganography and Digital Watermarking. Artech House,2000.
[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] Y. Cui and J. Widom, "Lineage Tracing for General Data Warehouse Transformations," The VLDB J., vol. 12, pp. 41-58,2003.
[8] F. Hartung and B. Girod, "Watermarking of Uncompressed and Compressed Video," Signal Processing, vol. 66, no. 3, pp. 283-301,1998.
[9] S. Jajodia, P. Samarati, M.L. Sapino, and V.S. Subrahmanian,"Flexible Support for Multiple Access Control Policies," ACM Trans. Database Systems, vol. 26, no. 2, pp. 214-260, 2001.
[10] Y. Li, V. Swarup, and S. Jajodia, "Fingerprinting Relational Databases: Schemes and Specialties," IEEE Trans. Dependable and Secure Computing, vol. 2, no. 1, pp. 34-45, Jan.-Mar. 2005.
[11] B. Mungamuru and H. Garcia-Molina, "Privacy, Preservation and Performance: The 3 P's of Distributed Data Management,"
technical report, Stanford Univ., 2008.
[12] V.N. Murty, "Counting the Integer Solutions of a Linear Equation with Unit Coefficients," Math. Magazine, vol. 54, no. 2, pp. 79-81,1981.
[13] S.U. Nabar, B. Marthi, K. Kenthapadi, N. Mishra, and R. Motwani, "Towards Robustness in Query Auditing," Proc. 32nd Int'l Conf. Very Large Data Bases (VLDB '06), VLDB Endowment, pp. 151-162,2006.
[14] P. Papadimitriou and H. Garcia-Molina, "Data Leakage Detection,"technical report, Stanford Univ., 2008.
[15] P.M. Pardalos and S.A. Vavasis, "Quadratic Programming with One Negative Eigenvalue Is NP-Hard," J. Global Optimization,vol. 1, no. 1, pp. 15-22, 1991.
[16] J.J.K.O. Ruanaidh, W.J. Dowling, and F.M. Boland, "WatermarkingDigital Images for Copyright Protection," IEE Proc. Vision, Signal and Image Processing, vol. 143, no. 4, pp. 250-256, 1996.
[17] R. Sion, M. Atallah, and S. Prabhakar, "Rights Protection for Relational Data," Proc. ACM SIGMOD, pp. 98-109, 2003.
[18] L. Sweeney, "Achieving K-Anonymity Privacy Protection Using Generalization and Suppression," http://en.scientificcommons.org/43196131, 2002.