**DRONACHARYA COLLEGE OF ENGIEERING**

**FARRUKH NAGAR, GURGAON**

**Synopsis Format**

1. **Name of Student:** Aayush Tiwari **Branch:-** CSIT **Roll no.** 19841
2. **Title :** Data Leakage Detection Using Cloud
3. **Objective/ Aim** : To propose data allocation strategies (across the agents) that improve the probability of identifying leakages of data’s and to look forward to know where the data’s are used by the third party.
4. **Literature Survey**: In the course of doing business, sometimes sensitive data must be handed over to supposedly trusted third parties. For example, a hospital may give patient records to researchers who will devise new treatments. Similarly, a company may have partnerships with other companies that require sharing customer data. Another enterprise may outsource its data processing, so data must be given to various other companies. We call the owner of the data the *distributor* and the supposedly trusted third parties the *agents*. Our goal is to *detect* when the distributor’s sensitive data has been *leaked* by agents, and if possible to identify the agent that leaked the data.

We consider applications where the original sensitive data cannot be perturbed. Perturbation is a very useful technique where the data is modified and made “less sensitive” before being handed to agents. For example, one can add random noise to certain attributes, or one can replace exact values by ranges [18]. However, in some cases it is important not to alter the original distributor’s data. For example, if an outsourcer is doing our payroll, he must have the exact salary and customer identification numbers. If medical researchers will be treating patients (as opposed to simply computing statistics), they may need accurate data for the patients.

Traditionally, leakage detection is handled by *watermarking*, e.g., 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. Watermarks can be very useful in some cases, but again, involve some modification of the original data. Furthermore, watermarks can sometimes be destroyed if the data recipient is malicious.

In this paper we study *unobtrusive* techniques for detecting leakage of a *set* of objects or records. Specifically, we study the following scenario: After giving a set of objects to agents, the distributor discovers some of those same objects in an unauthorized place. (For example, the data may be found on a web site, or may be obtained through a legal discovery process.) At this point the distributor can assess the likelihood that the leaked data came from one or more agents, as opposed to having been independently gathered by other means. Using an analogy with cookies stolen from a cookie jar, if we catch Freddie with a single cookie, he can argue that a friend gave him the cookie. But if we catch Freddie with 5 cookies, it will be much harder for him to argue that his hands were not in the cookie jar. If the distributor sees “enough evidence” that an agent leaked data, he may stop doing business with him, or may initiate legal proceedings.

For further instances we look back towards the major issues related to the data leakage scandals:-

1. First of all, we all use social networking sites like Facebook. In Facebook application, we have to give all the details related to the user like name, date of birth, educational qualification, pictures etc. We are promised by the Facebook that the data given by us is safe but is get stored in their system. In recent past times, there was a news that stunned everyone that the data available on the Facebook are leaked and the data are not secure enough to use. All the details including username, password, and personal pictures get leaked from the Facebook which was a big concern of all time. So there is a need of virtual system which can help in detecting the data leak so that the user can protect their data’s.
2. One of the major incidents of data leakage was done by EBAY, in which the username and password of more than 1 billion peoples got leaked along with their uploaded data’s.
3. In India, Aadhar Card is one of the most important documents used by us in daily life. It is hard to believe that even its data got leaked as Aadhar card has the retina scan and fingerprints of the individuals and as per CII of USA give the indication that its data can also get leaked.

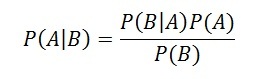
So nothing in the cloud which is used by the third party is safe enough and its data can be leaked easily. Therefore we need a better virtual machine which can provide a better environment for the user to deal online with the agents or third party.

1. **Methodology/ Planning of work**:

Explicit Data Requests:- Bays theorem of probability

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 *S* = {*t*1*,t*2} and there are two agents with explicit data requests such that *R*1 = {*t*1*,t*2} and *R*2 = {*t*1}. The value of the sum-objective is in this case:

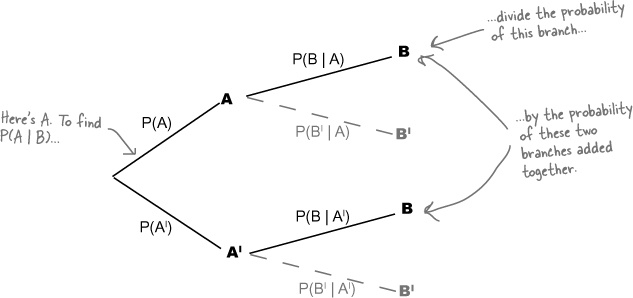


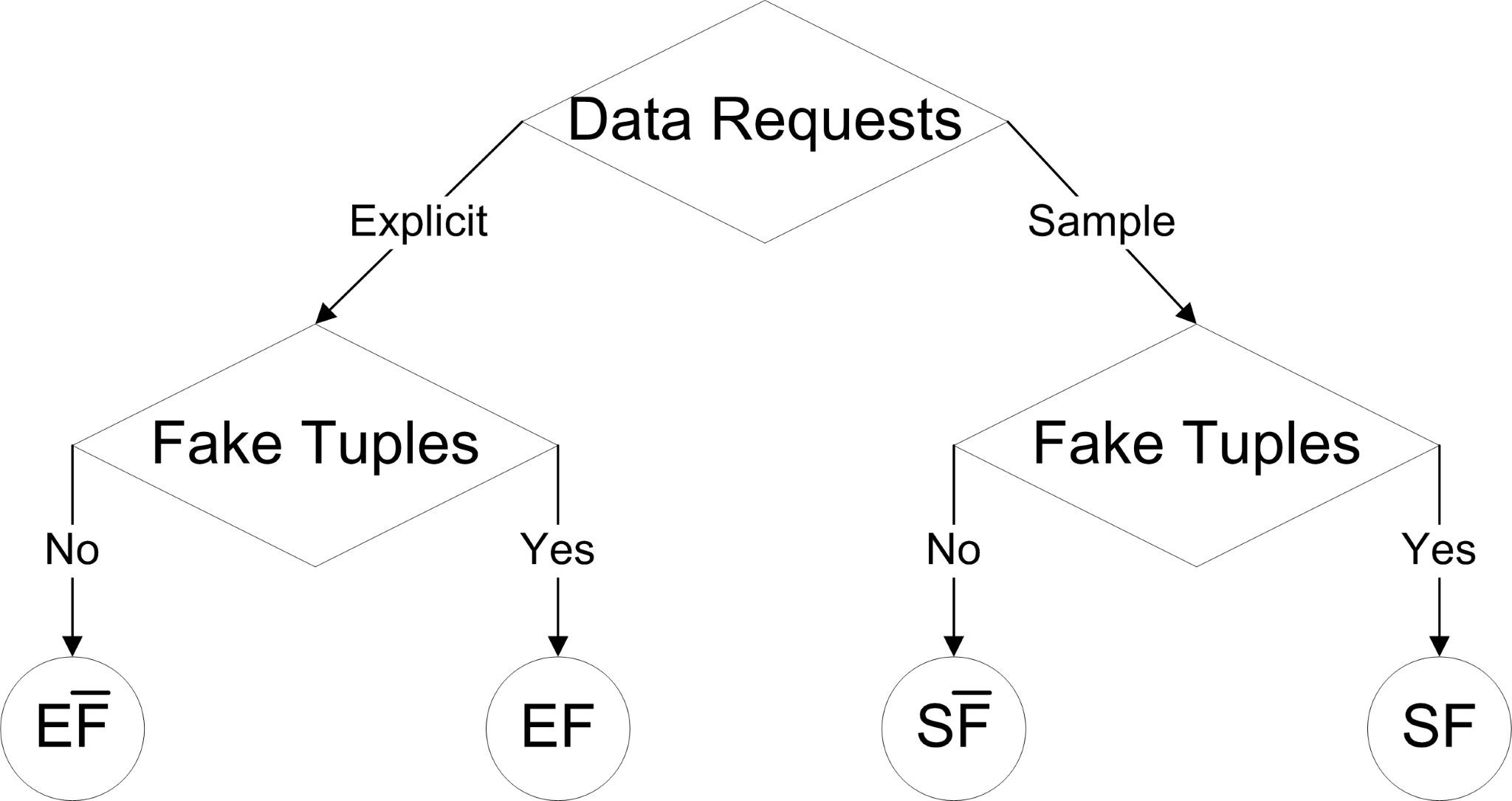
The distributor cannot remove or alter the *R*1 or *R*2 data to decrease the overlap *R*1 ∩ *R*2. However, say the distributor can create one fake object (*B* = 1) and both agents can receive one fake object (*b*1 = *b*2 = 1). In this case, the distributor can add one fake object to either *R*1 or *R*2 to increase the corresponding denominator of the summation term. Assume that the distributor creates a fake object *f* and he gives it to agent *R*1. Agent *U*1 has now *R*1 = {*t*1*,t*2*,f*} and *F*1 = {*f*} and the value of the sum-objective decreases to:

1-1/3=0.66*.*

bayes-4a.png (426Ã63)

1. **Facility required**(Hardware/ software) :-Azure Virtual Machine, Python Idle
2. **Technical details**:
3. Introduction: - Basically the project is all related to the combination of artificial intelligence and the cloud computing system in which we are going to develop a virtual cloud to store all the information’s released by the user and they can check that whatever the information is given to the third party is used at which domains or on which purpose.
4. Specifications: - The main specification of the system is that it can store data’s with the help of cloud computing virtually and artificial intelligence will come into work after the data transfer has taken place. After the data is transferred by the user it will get stored on the cloud for future reference and the data’s along with the information provided to the third party will be notified to the user and certainly whatever the third party do with that data will be all notified to the user. Any kind of leakage of data or misuse of data will alert the user about that particular information such that proper action can be taken and further there will be record of data given to the third party stored in the cloud can be used as a reference by user to get all the essential data back from the third party and so on. This is basically how the entire virtual system made by me will work with help of the artificial intelligence for providing a better and safe environment to the user for working online.
5. Block diagram/ Circuit diagram/ Flowchart:-



1. Control flow example:-
2. Brief description 
3. **Innovativeness and Usefulness: -** We are going to find a new and better way to safeguard the data and other personal information which is provided by the clients to the third party and some data’s goes missing and some get shared and can found at the unauthorized place. So to solve the uncertainty related to the misuse of the data given to the third party and to look after the the individual data’s, this software will help them to get rid of all the problems in very quick succession.
4. **Current status and Development:-** Currently the project in the research and development department where we are analyzing different types of leakages and the problems due to which it is being leaked. Is it a game of hackers or the technical fault of the third party system whatever it is we are going to solve that. We will certainly build a data leakage detection system which will help in ensuring a better experience to the population regarding the online transfer of data with the full information regarding their datas that where it is used.
5. **Market potential and competitive advantages :-** Talking about the market potential and advantages, we all know that how important and private data’s are for everyone and if any third party leaks it , it become a curse for them that why they have stored them. For solving their problem the data leakage detection technique will help them in an important manner. They will get each and every information regarding their data and if they provide any data to any third party they can track their data easily that where they are being used and for what purpose and if they find something irrelevant they can contact the third party about the same.
6. **Conclusion :-** When we are working online and providing data’s to third party it is not secure enough and may gets leaked. So for solving this problem a virtual machine is going to be made with cloud computing and with the help of algorithm of the artificial intelligence which will look after the data’s usage and it will notify user that where there data is being used and alert the user about the data leakage.

Overall the virtual machine will provide a secure and better environment for the user to work online along with the transfer of data without any problem of data leakage.

1. **References:** (1)Y . Li, V. Swarup, and S. Jajodia. Fingerprinting relational databases: Schemes and specialties. *IEEE Transactions on Dependable and Secure Computing*, 02(1):34–45, 2005.

(2) B. Mungamuru and H. Garcia-Molina. Privacy, preservation and performance: The 3 p’s of distributed data management. Technical report, Stanford University, 2008.

(3) V. N. Murty. Counting the integer solutions of a conditional probability with unit coefficients. *Mathematics Magazine*, 54(2):79–81, 1981.