Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 125 (2018) 662–668

Procedia Computer Science

www.elsevier.com/locate/procedia

6th International Conference on Smart Computing and Communications, ICSCC 2017, 7-8 December 2017, Kurukshetra, India

A Probabilistic Approach for Guilty Agent Detection using Bigraph after Distribution of Sample Data

Ishu Gupta*, Ashutosh Kumar Singh

Department of Computer Applications, National Institute of Technology Kurukshetra- 136119, India

Abstract

In today's on-growing world, there is need to share the information within or outside the enterprise which includes the sensitive information of the enterprise also. For example: companies has to share its sensitive information with its partners, employees and various other entities. This sensitive data can be leaked by third party. Later on, distributor finds the leaked documents at some unauthorized place (eg. through legal discovery process, on user's drive or the web). We propose a model which assesses the likelihood that the data has been leaked by one or more agents or it has been independently gathered by some other means. The goal of the model is to protect the sensitive information by detecting the leakage and identifying the leaker responsible for data leakage.

Peer-review under responsibility of the scientific committee of the 6th International Conference on Smart Computing and Communications

Keywords:
Bigraph, Confidential Information, Data Leakage, Fake Objects, Guilt Model, Sample Data Request, Security.

1. Introduction

The faster growth of information technology has resulted in copying and transferring of data in a ease manner [1], [2]. In the course of doing business, it is necessary to distribute the sensitive sample data to third party [3]. For example: a hospital can provide the patient's sample data to the various researchers [4], [5].

As the confidential data is shared among multiple entities; some of the entities can leak the data to the unauthorized third party which can be misused by it [6], [7]. Data leakage is the intentional or unintentional unauthorized transmission of the sensitive or private information to the malicious third party [8], [9]. The sensitive or private data in the enterprise includes various information such as intellectual property (IP), personal information (like account number, UID number, credit card data etc.), financial information, patient information etc. [10], [11], [12].

* Corresponding author.
E-mail address: ishugupta23@gmail.com, ashutosh@nitkkr.ac.in

1877-0509
Peer-review under responsibility of the scientific committee of the 6th International Conference on Smart Computing and Communications
10.1016/j.procs.2017.12.085

Data leakage has become a major problem being faced by the various enterprises and it is increasing day by day [13], [14], [15]. The loss of sensitive data causes the grievous threats to the organizations [16], [17]. It can decrease the goodwill and the reputation of the organizations [18]. Furthermore limiting the access to the sensitive information results in degradation in the organization's performance [19], [20].

There is a need of mechanism that can protect the sensitive information by detecting the leakage. Traditionally, watermarking is used to identify the leaker. The information is hide in each distributed copy of the document. On receiving the leaked document, hidden information is extracted and the guilty entity is find out but this phenomenon may involves modification or the complete distortion in the hidden information.

To solve the data leakage problem, we propose a guilty agent identification model that detects the leakage and identify the leaker who leaks the data. The model distributes the data among third party in such a manner which makes easy to identify the leaker. Furthermore, the model introduces the fake objects to identify the leaker with stronger evidence. Data allocation and distribution among the agents is represented through Bigraph. The benefits of the proposed approach is that it identifies the leaker on the basis of the data allocated to the various agents; hence does not involve any modification to the original data. Furthermore, the added fake objects acts as a type of watermark without engaging any alteration to it.

The rest of the paper is organized as follows. Section 2 describes the definition and mathematical background requires to understand the approach. Section 3 explains the proposed Guilty Agent Detection Model. Experimental results are shown in section 4. Finally, conclusion and future work are presented in section 5.

2. Definitions and Mathematical Backgrouond

2.1. Entities and Agents.

Distributor owns the sensitive data $\mathcal{D} = \{D_1, D_2, \ldots, D_n\}$ to be distributed among the various agents that are represented by the set $\mathcal{A}$ such that $\mathcal{A} = \{A_1, A_2, \ldots, A_m\}$ but does not want the data to be leaked to some malicious entity. An Agent A_j receives a set of objects $X_j \subseteq \mathcal{D}$ by requesting the sample data from the distributor. Sample data request is denoted as $SAMPLE(\mathcal{D}, r_j)$. In this request, a subset X_j consisting any r_j records from set $\mathcal{D}$ is provided to the agent A_j.

2.2. Guilty Agent.

The entity is said to be Guilty Agent G_A that leaks the data from its allocated dataset X_j to some unauthorized recipient which may cause misuse of data. Suppose after distributing the data to the agents, distributor finds that a set $\mathcal{L} \subseteq \mathcal{D}$ has been leaked. This means that it discovers some third party possessing $\mathcal{L}$ which is leaked by any guilty agent. Thus, an agent is said to be guilty who shares one or more objects with the third party.

2.3. Fake Objects.

The objects which appears realistic as the original data objects are consider as fake objects. Fake data objects denoted as F_j where $j = 1, 2, \ldots, m$ are distributed among agents with the sample data and it becomes typical for the agent to distinguish between the original object D_i and fake data object F_j. These objects are created with the intention of improving the chances of identifying G_A.

2.4. Bigraph.

A Graph $G(U, V, E)$ is said to be Bigraph whose vertices can be divided into two disjoint sets U and V such that $U \cap V = \phi$ and E is the set of edge from set U to set V. If $u', u'' \in U$ then $E(u', u'') \in \phi$ and $v', v'' \in V$ then $E(v', v'') \in \phi$. We can say that there can't exist any edge e between two vertices of the same set.