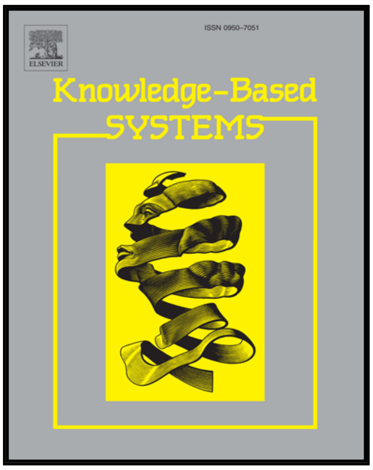

Data Leakage Detection Algorithm based on Task Sequences and Probabilities

César Guevara , Matilde Santos , Victoria López

PII: S0950-7051(17)30010-2
DOI: 10.1016/j.knosys.2017.01.009
Reference: KNOSYS 3786

To appear in: *Knowledge-Based Systems*

Received date: 11 May 2016
Revised date: 4 January 2017
Accepted date: 5 January 2017

Please cite this article as: César Guevara , Matilde Santos , Victoria López , Data Leakage Detection Algorithm based on Task Sequences and Probabilities, *Knowledge-Based Systems* (2017), doi: 10.1016/j.knosys.2017.01.009

Data Leakage Detection Algorithm based on Task Sequences and Probabilities

Authors: César Guevara, Matilde Santos, Victoria López*

Affiliation: Computer Science Faculty, C/ Profesor García Santesmases, s/n, Complutense University of Madrid, 28040-Madrid, Spain

E-mail: {cesargue, msantos, vlopezlo}@ucm.es

* Corresponding author phone: +34 91 394 76 20

Abstract. In this paper we propose a novel algorithm to detect anomalous user behaviour in computer sessions. We first identify the behavioural profile of each authorized user from the computational tasks they usually carry out on the files of the information system. A new session is then codified as 2-length sequences and an algorithm based on the probability of those sequences is applied. The activities classified as possible anomalies are double-checked by applying Markov chains. The procedure has been proved efficient in terms of high detection accuracy and low false positive rate. It has been validate on a real database provided by a governmental institution of Ecuador and also on a public dataset of Unix commands. Besides, the algorithm has been shown efficient regarding computational time and the overhead of this monitoring software is low.

Keywords: Anomaly detection, Computer user behaviour, Markov chains, Data leakage, Knowledge-based decision system.

1. Introduction

Data Leakage (DL) is a serious and increasingly common problem for organizations, companies and institutions around the world. This problem can be defined as the unauthorized transfer of classified information from a computer or data centre to the outside world. That is, theft of computer based sensitive information. Sensitive data in companies and organizations include intellectual property, financial information, patient information, personal credit-card data, and any other information depending on the business and the industry involved (Lewellen et al., 2012; Huth et al. 2013).

The study presented in Cisco (2008) reveals that organizations experienced about one threat per month. A significant proportion of computer and organizational security professionals believe insider threat is the greatest risk to their enterprise. Even more, internal fraud produces bigger financial losses in comparison to