Journal of Network and Computer Applications 62 (2016) 137–152

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

A survey on data leakage prevention systems

Sultan Alneyadi *, Elankayer Sithirasenan, Vallipuram Muthukkumarasamy

School of Information and Communication Technology, Griffith Sciences, Griffith University, Gold Coast, Queensland 4222, Australia

ARTICLE INFO

Article history:
Received 22 May 2014
Received in revised form
14 November 2015
Accepted 6 January 2016
Available online 19 January 2016

Keywords:
Data leakage detection
Data leakage prevention
Data security
DLPSs
Leaking channels
Content analysis

ABSTRACT

Protection of confidential data from being leaked to the public is a growing concern among organisations and individuals. Traditionally, confidentiality of data has been preserved using security procedures such as information security policies along with conventional security mechanisms such as firewalls, virtual private networks and intrusion detection systems. Unfortunately, these mechanisms lack pro-activeness and dedication towards protecting confidential data, and in most cases, they require predefined rules by which protection actions are taken. This can result in serious consequences, as confidential data can appear in different forms in different leaking channels. Therefore, there has been an urge to mitigate these drawbacks using more efficient mechanisms. Recently, data leakage prevention systems (DLPSs) have been introduced as dedicated mechanisms to detect and prevent the leakage of confidential data in use, in transit and at rest. DLPSs use different techniques to analyse the content and the context of confidential data to detect or prevent the leakage. Although DLPSs are increasingly being designed and developed as standalone products by IT security vendors and researchers, the term still ambiguous. In this study, we have carried out a comprehensive survey on the current DLPS mechanisms. We explicitly define DLPS and categorise active research directions in this field. In addition, we suggest future directions towards developing more consistent DLPSs that can overcome some of the weaknesses of the current ones. This survey is an updated reference on DLPSs, that can benefit both academics and professionals.

1. Introduction

Prevention of data disclosure to unauthorised entities is one of the main goals in information security. It continuously and rapidly drives both academic and industrial sectors to investigate, design and develop different security solutions to mitigate the risk of data leakage. However, preventing data leakage is not always possible because of the need to access, share and use information, which leads to inevitable release of confidential data. This revelation comes in the form of information leak, which might be the result of a deliberate action or a spontaneous mistake. Recent reports indicate growing concerns in government and business sectors as a result of data leakage. According to datalossdb (2015), in year 2014, about 50% of recorded data leakage occurred in the business sector, about 20% occurred in the government sector and about 30% occurred in the health and education sectors. Private users are also affected from data leakage, but it is hard to know the exact amount and severity of private data leakage. Although some reported leaks were not detrimental to organisations, others have caused several million dollars' worth damage. Business credibility is compromised when sensitive data such as future projects, trade secrets and customer profiles are leaked to competitors. Government data leaks may involve sensitive information about political relationships, law enforcement and internal security. A popular incident involving leaked sensitive government information was the release of the United States diplomatic cables by WikiLeaks. The leak consisted of about 250,000 United States diplomatic cables and 400,000 military reports referred to as 'war logs'. This revelation was carried out by an internal entity using an external hard drive and about 100,000 diplomatic cables were labelled *confidential* and 15,000 cables were classified as *secret* (Karhula, 2011). This incident received a high level of attention as the United States faced much criticism from governments and civil rights organisations worldwide. Another famous incident was the release of 77 million account details of Sony PlayStation network subscribers (Arthur and Stuart, 2011). The leak was due to an external intrusion, which forced the PlayStation network services to shut down for more than 24 days. This incident seriously impacted the reputation of Sony, receiving much criticism from users, and eventually led to a public apology from Sony's chief executive officer. One of the biggest recorded data leakage incidents was the release of names, email addresses and personal data of eBay customers (Wakefield, 2014), where around 145 million customers

* Corresponding author. Tel.: +64 24586815.
E-mail addresses: sultan.alneyadi2@griffithuni.edu.au (S. Alneyadi), e.sithirasenan@griffith.edu.au (E. Sithirasenan), v.muthu@griffith.edu.au (V. Muthukkumarasamy).

http://dx.doi.org/10.1016/j.jnca.2016.01.008
1084-8045/

were affected severely disrupting the business. These kinds of incidents can cause major financial losses and severely damage an organisation's reputation.

Driven by the need to address such serious issues, security experts endeavour to develop various security measures. Systems such as firewalls, intrusion detection systems (IDSs) or intrusion prevention systems (IPSs), and virtual private networks (VPNs) have been introduced over the past three decades. These proven systems can perform satisfactorily if the data to be protected is well defined, structured and constant. However, using these measures to protect evolving (i.e. edited, differently tagged or compressed) confidential data can be naive. For example, a firewall can block access to a confidential data segment using simple centralised rules; however, the same data segment may be accessible through other means such as an email attachment or instant messaging (IM). Thus, conventional security measures (i.e. firewalls, IDSs, VPNs) lack persistency and understanding of data semantics. To overcome this deficiency, a new direction for data protection was considered leading to the introduction of data leakage (loss) prevention systems (DLPSs). DLPSs are especially designed systems that have the ability to identify, monitor and protect confidential data and detect misuse based on predefined rules. The DLP field is considered relatively new compared with conventional security solutions. Moreover, to many academics and security practitioners, the field is indistinguishable because adequate research, surveys or both are lacking at present.

Motivated by the significance of the DLP field of study and the need for better understanding of current and future DLP trends, we present this survey paper. This paper contributes to the DLP field by explaining the DLP paradigm, including data states and deployments. Further, it identifies the challenges facing DLPSs. Moreover, it comprehensively gathers, categorises, discusses and compares the current DLP methods in industry and academia. It also lists and discusses DLP analysis techniques; and presents future DLPS trends.

This paper is structured as follows. Section 2 discusses the DLP paradigm. Section 3 describes the challenges facing DLPSs. Section 4 categorises the current DLP methods and discusses the advantages and disadvantages of each method. Section 5 explains the DLPS analysis techniques. Section 6 suggests future DLPS trends. Section 7 discusses the survey limitations. Section 8 concludes the survey paper.

2. Data leakage prevention

A number of attempts to study and define the area of data leakage prevention have been made in both academia and industry. These attempts discuss DLPSs from differing perspectives because DLPSs are still new and there is no concrete agreement on a common definition yet. Both academics and practitioners are using various names for DLPSs, such as data loss/leak prevention, information loss/leak prevention, extrusion prevention and content monitoring and filtering/protection (Mogull, 2010).

In academia, some researchers have provided a broad idea about the DLP research area. For example, a review paper by Raman et al. (2011) discussed the importance of the DLP research area and suggested that more attention be paid to it. The authors mentioned common DLP approaches and associated problems. In addition, they suggested new directions for future work, and introduced text clustering and social network analysis as future solutions for the problem. A more comprehensive survey on DLP was presented by Shabtai et al. (2012). The authors define a DLPS as 'a system that is designed to detect and prevent the unauthorised access, use, or transmission of confidential information' (p. 10). Their survey describes taxonomy of data leakage prevention solutions along with commercial and academic examples. Academic DLP methods are categorised into misuse detection in information retrieval systems/database, email protection, network/web-based protections, encryption and access control, data hidden in files and honeypots/honeytokens (p. 22). Data leakage/misuse scenarios, case studies and future trends are also given in this survey.

Professional and industrial institutes have also put effort into addressing the DLP area, including SANS, Securosis and ISACA. SANS presented a white paper (Kanagasingham, 2008) that provides a brief history about DLP solutions and how they fit within other network security technologies. Mogull (2010) from Securosis presented a white paper on understanding and selecting a DLP solution. The paper discusses the DLP market, in general, and the difference between a DLP feature and a DLP solution. It also considers the confusion surrounding the definition of DLPS and the variation in commercial products among vendors, which has resulted in the same product having many different names. Mogull defines DLPSs as 'products that, based on central policies, identify, monitor, and protect data at rest, in motion, and in use, through deep content analysis' (p. 5) and explains the differences between content and context analysis, suggesting that the former is more promising than the latter. Finally, the paper provides a summary of the strengths and weaknesses of the current content analysis approaches, such as rule-based or regular expressions, fingerprinting, exact file matching, partial document matching and statistical analysis.

ISACA's (2010) white paper discusses DLPSs from a management point of view. It suggests that implementing a DLPS must be thoroughly planned and studied in terms of the need, the size and the aim of the organisation. The paper explains that unplanned implementation can defeat the purpose of using a DLPS in the first place. For example, if an organisation is using a DLPS to avoid business loss, business can be disrupted by wrong implementation of a DLPS. Wrong implementation includes hindering workflow by extensive traffic inspection and weak integration with other security mechanisms. The paper also discusses the many challenges that must be addressed before using a DLPS to ensure an organisation is ready to use it. The specific challenges vary among organisations, depending on the nature of the business and the volume of transactions.

2.1. Data leakage prevention systems

Data leakage (or data loss) is a term used in the information security field to describe unwanted disclosures of information. This problem is mitigated by using different DLP methods and techniques, including both administrative and technical approaches. In this paper, we define DLPSs as designated analytical systems used to protect data from unauthorised disclosure at all states using remedial actions triggered by a set of rules. This definition contains three main attributes that distinguish DLPSs from conventional security measures. First, DLPSs have the ability to analyse the content of confidential data and the surrounding context. Second, DLPSs can be deployed to provide protection of confidential data in different states, that is, in transit, in use and at rest. The third attribute is the ability to protect data through various remedial actions, such as notifying, auditing, blocking, encrypting and quarantining. The protection normally starts with the ability to detect potential leaks through heuristics, rules, patterns and fingerprints. The prevention then happens accordingly.

DLPSs differ from conventional security controls such as firewalls, VPNs and IDSs in terms of dedication and proactivity. Conventional security controls have less dedication towards the actual content of the data. They might block users' access to data for the sake of sensitive data protection in the case of firewalls, or simply encrypt all the traffic, as in the case of VPNs, which might include