Understanding Data Leak Prevention

Preeti Raman, Hilmi Güneş Kayacık, and Anil Somayaji

***Abstract*—Data leaks involve the release of sensitive information to an untrusted third party, intentionally or otherwise. Many vendors currently offer data leak prevention products; surprisingly, however, there is very little academic research on this problem. In this paper, we attempt to motivate future work in this area through a review of the field and related research questions. Specifically, we define the data leak prevention problem, describe current approaches, and outline potential research directions in the field. As part of this discussion, we explore the idea that while intrusion detection techniques may be applicable to many aspects of the data leak prevention problem, the problem is distinct enough that it requires its own solutions.**

***Index Terms*—Data Leak Prevention, Text Clustering Analysis, Social Network Analysis**

I. Introduction

ORGANIZATIONS increasingly may be harmed by data being revealed to unauthorized parties. Such data leaks can cause harm in a variety of ways. Improper handling of confidential data can violate government regulations, resulting in fines and other sanctions. Companies can be held liable for the release of customer and employee information such as credit cards and social security numbers. Further, loss of proprietary information to competitors can result in loss of sales and may even threaten the existence of an organization.

Data leak prevention (DLP) refers to products or techniques that attempt to mitigate some or all of these threats. DLP products are available from multiple vendors, including Symantec [1], CA Technologies [2], Trend Micro [3] and McAfee [4]. In contrast, data leak prevention has received little attention in the academic research community. This is not to say that DLP is a solved problem: indeed, current products are limited in what threats they address.

In this paper, we argue that data leak prevention is an area ripe for further research in that there are multiple hard problems of significant real-world interest that have not been rigorously studied. While DLP overlaps significantly with the field of intrusion detection, as we will explain that the overlap in potentially applicable techniques obscures the more significant differences in the respective problem requirements.

The rest of this paper proceeds as follows. First, we define the data leak prevention problem in Section II. Section III describes past related work in DLP. We describe the challenges of the problem in Section IV. Section V presents potential research directions in DLP. We address the issue of the overlap between DLP and intrusion detection and conclude in Section VI.

II. Data Leak Prevention Problem

There are numerous ways sensitive data can be revealed to untrusted third parties, as depicted in Figure 1. Thus, in order to discuss the data leak prevention problem, we investigate several factors including data repositories and available data leak channels. It is crucial to identify sensitive data repositories within an organization since selecting suitable prevention techniques naturally depends on the repository in question. Customer records, proprietary source code and sensitive documents on network shares are a few examples of repositories. Different prevention techniques may be appropriate for different data states: 1) at rest (at the repository); 2) in motion (over the network), and 3) in use (at the endpoint) [5]. When the data is at rest, the repository can be protected with access control and audit. However, when the data is in motion or in use, prevention using access control becomes increasingly difficult. For in motion and in use scenarios, the data leak prevention mechanism should be sufficiently context aware to infer the semantics of communication.

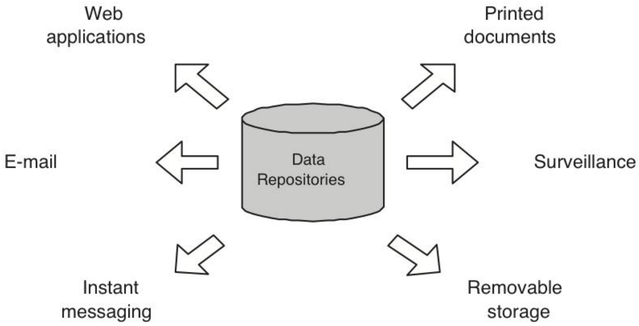

Fig. 1. Data leak channels.

As shown in Figure 1, data leaks can occur in many different ways. Hardware theft, social engineering,

The authors gratefully acknowledge the support of CA Labs, MITACS-Accelerate, NSERC Discovery and NSERC ISSNet funding programs.

P. Raman, H. G. Kayacık, and A. Somayaji are in the School of Computer Sicnece, Carleton University, Ottawa, ON, K1S 5B6, Canada (email: {praman, kayacik, soma}@ccsl.carleton.ca}

surveillance, and the mismanagement of printed documents are a few of the more traditional data leak channels. Additionally, electronic communications such as instant messaging, web applications and email provide additional challenges. These electronic channels are highly utilized in organizations and provide means to quickly and easily send data to a third party. While traditional data leaks can be more suitably defended with traditional approaches [6], lightweight and context aware techniques, which can infer who is communicating and what is being communicated, are needed to prevent data leaks in electronic communications.

III. Current Approaches

Various companies have recently started providing data leak prevention solutions. While some solutions secure 'data at rest' by restricting access to it and encrypting it, the state of the art relies on robust policies and pattern-matching algorithms for data leak detection. On the other hand, related academic work in data leak prevention focused on building policies [7], developing watermarking schemes [8], and identifying the forensic evidence for post-mortem analysis [9].

Vachharajani et al. [7] provides a user-level policy language for hardware-enforced policies, which ensures that the sensitive data does not reach untrusted output channels through network communications, files, and shared memory. The proposed runtime information flow security system assigns predefined labels to the data and policies are enforced at the hardware level to ensure the data flow complies with the policies. Needless to say, such an approach involves the labor-intensive task of the definition of labels, policies, and requires hardware that supports information flow security.

Lee et al. [9] approaches data leak prevention from a forensics point of view and identifies the set of files needed to detect data leaks on a Windows operating system. The authors argue that delaying the collection of forensic data will have detrimental effects in the effectiveness of a data leak prevention system; hence, they propose an efficient method to collect the basic information needed to detect data leaks by investigating five crucial system files: 1) the installation record file, 2) the system event log, 3) the windows registry, 4) the browser history, and 5) the core file in NTFS. Their approach is limited to file system-level data leaks on Windows platforms.

The synthetic decoy scheme of White et al. [8] focuses on the data leaks on large databases of personal records and proposes realistic decoy records to identify the source of data leaks, particularly when multiple databases are concerned. By creating uniquely identifiable, but semantically plausible personal records, the database can be digitally watermarked. Thus, any data leak from the database will contain the decoys unique to the database in question, hence, revealing the source of the leak. Such an approach, by nature, focuses on the postmortem identification of the data leak source not the real-time detection of the leak itself.

The current state-of-the-art in commercial data leak prevention focuses on pattern-matching, which suffers from the general shortcoming of misuse detection techniques; an expert needs to define the signatures. Given the elusive definition of data leaks, signatures should be defined per corporation basis, making the widespread deployment of current data leak prevention tools a challenge. On the other hand, the relevant academic work on data leak prevention and text mining takes a forensics approach and mainly focuses on post-mortem identification. Thus, detecting complex data leaks in real-time remains an understudied field.

IV. Challenges

A. Encryption

As discussed in Section II, different prevention mechanisms are needed to cover different states of data. In particular, detecting and preventing data leaks in transit are hampered due to encryption and the high volume of electronic communications. While encryption provides means to ensure the confidentiality, authenticity and integrity of the data, it also makes it difficult to identify the data leaks occurring over encrypted channels. Encrypted emails and file transfer protocols such as SFTP imply that complementary DLP mechanisms should be employed for greater coverage of leak channels. Employing data leak prevention at the endpoint – outside the encrypted channel – has the potential to detect the leaks before the communication is encrypted.

B. Access Control

Access control provides the first line of defense in DLP. However, it does not have the proper level of granularity and may be outdated. While access control is suitable for data at rest, it is difficult to implement for data in transit and in use. In other words, once the data is retrieved from the repository, it is difficult to enforce access control. Furthermore, access control systems are not always configured with the least privilege principle in mind. For example, if an access control system grants full access to all code repositories for all programmers, it will not effectively detect data leaks where a programmer accesses a project that he/she is not involved in.

C. Semantic Gap in DLP

DLP is a multifaceted problem. The definition of a data leak is likely to vary between organizations depending on the sensitive data to be protected, the degree of interaction between the users and the available communication channels. The current state-of-the-art, which is reviewed in Section III, mainly focuses on the use of misuse detection (signatures) and post-mortem analysis (forensics). The common shortcoming of such approaches is that they lack the semantics of the events being monitored. When a data leak is defined by the communicating parties as well as the data exchanged during the communication, a simple pattern matching or access control scheme cannot infer the nature of the communication. Therefore, data leak prevention mechanisms need to keep track of who, what and where to be able to defend against complex data leak scenarios.

TABLE I
A SUMMARY OF RELEVANT DATA LEAK PREVENTION MECHANISMS

	Data State	Data Channel	Detection based on	Objectives and Remarks
Pattern matching	In use	• E-communications	Database of data leak signatures	• Develop misuse signatures. • If signature match occurs, indicates data leak. • Attack mutations/modifications are hard to handle.
Access control	At rest	• Databases • Repositories	Access control list	• Control the access of a resource. • Grant access if user is on the “white-list.” • A perimeter defense, hence does not work proactively.
Text clustering	In transit In use	• E-communications • Repositories (focused on text)	A set of clusters with semantic meaning of he communications	• Identify the nature/topic of communication. • Unusual activity requires attention. • Needs to be scalable and results should be easy to comprehend.
Social network analysis	In transit In use	• E-communications • Repositories (focused on user interaction)	Social network graph of users	• Discover social networks of collaboration. • Drastic changes in social networks require attention. • Social networks need validation before use.

A. Collaboration

In order to be able to identify the ‘outsider’ in a communication, the collaborating parties should be identified. However, identifying collaborators is not a straightforward task. While a naive metadata approach can consider using access control mechanisms (e.g. to determine the programmers, managers, administrators, etc.) such an approach is not sufficient to capture heterogeneous groups where people can belong to more than one group. Furthermore, the temporal nature of collaborations should be addressed. As time passes, new collaborations are formed and existing ones disappear. Thus, the analysis of collaborations should not be regarded as a one-time task but as a continuous task to be carried out on regular intervals.

V. FUTURE DIRECTIONS

As summarized in Table I, the biggest shortcoming of the state-of-the-art and the relevant previous work is that they attempt to detect data leaks without an understanding of the communication context. However, the complex data leaks are in semantics (i.e. the content of the conversation) not in syntax (i.e. whether a pattern resembling social insurance numbers occurs). Thus, in order to address the semantic gap problem in data leak prevention, new research directions should be explored to provide the semantic summarization of communications. The main focus is identifying in-transit and in-use data leaks. In this section, we review the text clustering and social network analysis approaches that are likely to aid in building context aware DLP solutions.

A. Text Clustering

Text clustering [10] is an exploratory data analysis technique that aims to identify the natural groupings (i.e. ‘clusters’) within a text corpus. Each cluster contains similar documents, according to a similarity metric such as Euclidean distance. From a data leak prevention perspective, text can be collected from numerous sources, an example of which is email. The clusters of text can serve as equivalence classes (content summaries) that can then be labeled to provide semantic meaning. Thus, by applying clustering to email communications, it is possible to infer the subject of the communication in a privacy-preserving manner. Based on the subjects about which a user communicates, a deviation from the ‘usual’ is flagged and further analyzed for data leaks.

Text clustering, which places documents with similar properties within the same group, has been utilized for summarizing large corpus of documents. Cavnar et al. [11] employed n-gram representation of text-for-text categorization. The documents are represented as n-grams, in which an n-gram is an n character slice of a longer string. Taking advantage of the Zipf’s Law [12] in human language text, they identified the language of the text based on the most frequent 300 n-grams. Furthermore, they demonstrated that, the n-grams below 300 are specific to text topic, hence providing a means to cluster the text according to context.

In terms of the analysis of email as a text corpus, Chow et al. [13] aimed to detect the inferences in sensitive documents by applying various data mining algorithms to the Enron email corpus [17], which contains the email communications of top-level Enron employees before and during the Enron scandal. The inferences are determined based on co-occurrence of terms in the text corpus. Similarly, Keila et al. [14] proposed a method for detecting deceptive emails, based on the expectation that people use fewer first person pronouns and more negative emotion and action verbs. Singular value decomposition is utilized to visualize email messages and identify the outliers, which correspond to deceptive emails. The previous relevant text mining approaches [11][13][14] focus on document summarization in general, without a data leak prevention focus.

Applying text clustering to data leak prevention involves monitoring corporate email communications for a period of time to identify the clusters of topics, in other words, communication subjects. The output of clustering may be difficult for a human to comprehend without further processing such as in the case of the commonly utilized k-means clustering, which represents centroids (i.e. cluster centers) as high-dimensional vectors. However, clustering algorithms such as approximate divisive hierarchical clustering [10][15] can provide a cluster-identifying tree, which the administrator can analyze and modify, if necessary. Thus the resulting visualization can be utilized to assign semantic meaning to the clusters manually or automatically. During deployment, when an email communications is processed, the most similar cluster is employed to assign the topic of the email. If there exists a substantial deviation of communication pattern (in terms of the context, frequency and the involved parties), the resulting communication is flagged for further analysis.

B. Social Network Analysis

Social network analysis involves the mapping and measuring of relationships between people, groups, and organizations by representing the relationships in terms of nodes and connections. Social networks can be derived from communication channels such as email, forum discussions, and social networking sites. Analysis of social networks can improve our understanding of the relationships and groupings between the parties involved in electronic communications, email in particular. Thus, the goal of social network analysis for data leak prevention is to identify the communication patterns within the organization and employ feedback from the administrator to identify unusual communications to uncover to data leaks.

Diesner et al. [16] performed a social network analysis of the Enron emails. The social networks extracted from the email communications take the form of directed graphs where each edge is weighted according to the cumulative frequency of emails exchanged between the nodes (i.e. people) in the graph. The comparison of the communication structure before and during the crisis indicated a movement toward communicating only between trusted parties, due to accountability. Furthermore, immediately after the bankruptcy became public, an increase in outward communications is observed potentially a likely outcome of people seeking more information on the recent events [16].

Applying social network analysis in data leak prevention involves monitoring online collaboration (email, document and code repositories) to discover the social networks of collaboration. The discovered social networks are vital in identifying collaborators such as a team of developers working on the same code repository or a group of employees exchanging emails to perform a task (e.g. preparing for a meeting). Social network analysis has the potential to discover collaborations that are not documented as a part of company policy or access control. Proper visualization of social networks can be presented to the administrator for manual or automatic validation. During deployment, if a substantial change in the social network is observed, it is flagged for further analysis since it can reveal: 1) a dissolving social network, 2) a merging social network, or 3) inclusion of an untrusted party, which is potentially a data leak.

VI. Conclusion

DLP is a multifaceted problem. Determining the sensitive data to be protected, identifying the legitimate use of the data and anticipating data leak channels require knowledge of the internal business logic of the corporation. Thus, there is no one-size-fits-all solution. In addition to traditional data leak channels such as hardware theft, the widespread use of electronic communications such as email makes it easy to leak sensitive data in a matter of seconds. Both data leak prevention and intrusion detection share the same common goal, which is to detect potentially harmful activity. Thus, the commercial approach typically employs similar techniques to solve data leak prevention. However, data leak prevention focuses on what (is leaked) as opposed to intrusion detection, which focuses on who (is breaking in). DLP is a complex problem, in which the threat usually originates from the 'inside' and the definition of 'misuse' is elusive. Data leaks can occur by accident between individuals who are completely legitimate. The detection of such data leaks requires an understanding of semantics. The current state-of-the-art in data leak prevention mainly utilizes misuse detection to detect data leaks, where a signature acts as a data leak description. However, misuse detection cannot scale well in data leak prevention since the data leak signatures – highly dependent on the internal business logic – should be developed per organization to minimize false positives and maximize detection rate. Furthermore, misuse detection does not possess the sufficient context awareness to detect complex data leak scenarios, where the data leak is in the semantics, not in syntax.

In this paper, we reviewed the current state-of-the-art as well as potential research areas which can provide context-aware data leak prevention solutions, as summarized in Table I. Text clustering and social network analysis discussed in Section V focus on summarizing the electronic communications in a lightweight privacy-preserving manner and inferring the semantic meaning. This allows data leak prevention to go beyond pattern-matching and detect complex data leaks based on who is involved in the communication well as what information is being exchanged.

Acknowledgment

The authors would like to thank Carrie Gates from CA

Labs for the valuable feedback.

REFERENCES

[1] Machine Learning Sets New Standard for Data Loss Prevention: Describe, Fingerprint, Learn [White Paper]. (2010, December 14). *Symantec*. [Online]. Available: http://eval.symantec.com/mktginfo/enterprise/whitepapers/b-dlp machine learning.WP en-us.pdf

[2] CA DLP: information protection and control [Product Sheet]. (2010). *CA Technologies* [Online]. Available: http://www.ca.com/~/media/Files/productbriefs/dlp-12-5-ps.pdf

[3] Trend micro data protection: Solutions for privacy, disclosure and encryption [White Paper]. *Trend Micro* [Online]. Available: http://us.trendmicro.com/.../datalossprevention/wp02_dlp-compliance-solutions_100225us.pdf

[4] McAfee host data loss prevention [Data Sheet]. *McAfee* [Online]. Available: http://www.mcafee.com/us/resources/data-sheets/ds-host-data-loss-prevention.pdf

[5] Data leak prevention [White Paper]. (2010, September 14). *Information Systems Audit and Control Association (ISACA)* [Online]. Available: http://www.isaca.org/Knowledge-Center/Research/Documents/DLP-WP-14Sept2010-Research.pdf

[6] J. Livingston, "Tips and Strategies to Protect Laptops and the Sensitive Data They Contain," *Information Systems Audit and Control Association (ISACA) Journal*, vol. 5, pp. 1-3, 2007.

[7] N. Vachharajani, M. J. Bridges, J. Chang, R. Rangan, G. Ottoni, J. A. Blome, G. A. Reis, M. Vachharajani, and D. I. August, "Rifle: An architectural framework for user-centric information-flow security," in Proceedings of the 37th annual IEEE/ACM International Symposium on Microarchitecture (MICRO 37), Portland, OR, USA, 2004, pp. 243–254.

[8] J. White, and D. Thompson, "Using synthetic decoys to digitally watermark personally-identifying data and to promote data security," in Proceedings of the International Conference on Security and Management (SAM 2006), June 2006, pp. 91–99.

[9] S. Lee, K. Lee, A. Savoldi, and S. Lee, "Data leak analysis in a corporate environment," in Proceedings of the 2009 Fourth International Conference on Innovative Computing, Information and Control (ICICIC '09), Las Vegas, NV, June 2009, pp. 38–43.

[10] J. Han. *Data Mining: Concepts and Techniques.* San Francisco, CA, USA: Morgan Kaufmann Publishers, Inc., 2005.

[11] W. B. Cavnar, and J. M. Trenkle, "N-gram-based text categorization," In *Proceedings of 3rd Annual Symposium on Document Analysis and Information Retrieval (SDAIR-94),* 1994, pp. 161–175.

[12] G. K. Zipf, *Human Behavior and the Principle of Least-Effort.* Cambridge, MA: Addison-Wesley, 1949.

[13] R. Chow, P. Golle, and J. Staddon, "Detecting privacy leaks using corpus-based association rules," in *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08)*. Las Vegas, NV, USA, August 2008, pp. 893–901.

[14] P. S. Keila, and D. B. Skillicorn, "Detecting unusual email communication," in *Proceedings of the 2005 conference of the Centre for Advanced Studies on Collaborative research (CASCON '05),* IBM, October 2005, pp. 117–125.

[15] H. Inoue, D. Jansens, A. Hijazi, and A. Somayaji, "Netadhict: a tool for understanding network traffic," in *Proceedings of the 21st conference on Large Installation System Administration Conference (LISA'07),* Dallas, TX, USA, November 2007, pp. 1–9.

[16] J. Diesner, T. L. Frantz, and K. M. Carley, "Communication networks from the Enron email corpus: It's always about the people. Enron is no different," *Computational & Mathematical Organization Theory* [Online], vol. 11, October 2005, pp. 201–228. Available: http: //portal.acm.org/citation.cfm?id=1110938.1110942

[17] J. Shetty, and J. Adibi, "The Enron email dataset database schema and brief statistical report," Information Sciences Institute, University of Southern California, Los Angeles, CA, USA, Technical Report, 2004.