Overview

Enterprise data breach: causes, challenges, prevention, and future directions

Long Cheng, Fang Liu and Danfeng (Daphne) Yao*

A data breach is the intentional or inadvertent exposure of confidential information to unauthorized parties. In the digital era, data has become one of the most critical components of an enterprise. Data leakage poses serious threats to organizations, including significant reputational damage and financial losses. As the volume of data is growing exponentially and data breaches are happening more frequently than ever before, detecting and preventing data loss has become one of the most pressing security concerns for enterprises. Despite a plethora of research efforts on safeguarding sensitive information from being leaked, it remains an active research problem. This review helps interested readers to learn about enterprise data leak threats, recent data leak incidents, various state-of-the-art prevention and detection techniques, new challenges, and promising solutions and exciting opportunities.

How to cite this article:
WIREs Data Mining Knowl Discov 2017, e1211. doi: 10.1002/widm.1211

INTRODUCTION

Data leakage is a serious threat to enterprise operations, such as corporations and government agencies. The loss of sensitive information can lead to significant reputational damage and financial losses, and even can be detrimental to the long-term stability of an organization. Common types of leaked information range from employee/customer data, intellectual property, to medical records. According to IBM's 2016 Cost of Data Breach Study,[1] the average consolidated cost of a data breach has reached $4 million. Juniper Research's forecast[2] suggests that the global annual cost of data breaches will be over $2.1 trillion globally by 2019, due to the rapid digitization of consumers' lives and enterprise records. Over the past few years, there have been many notable data loss incidents that cost companies millions of dollars. Cybercriminals breached the Target Corporation's network in 2013, stealing 40 million payment cards information and 70 million customers' personally identifiable information, which has incurred $248 million losses to date reported by Target.[3] In 2016, Yahoo reported that at least 500 million accounts in 2014 had been stolen in an apparent 'state sponsored' data breach.[4] Since data volume is growing exponentially in the digital era and data leaks happen more frequently than ever before, preventing sensitive information from being leaked to unauthorized parties becomes one of the most pressing security concerns for enterprises.

Data leakage can be caused by internal and external information breaches, either intentionally (e.g., data theft by intruders or sabotage by insider attackers) or inadvertently (e.g., accidental disclosure of sensitive information by employees and partners). A study from Intel Security[5] showed that internal employees account for 43% of corporate data leakage, and half of these leaks are accidental. Motivations of insider attacks are varied, including

*Correspondence to: danfeng@vt.edu

Department of Computer Science, Virginia Tech, Blacksburg, VA, USA

Conflict of interest: The authors have declared no conflicts of interest for this article.

corporate espionage, grievance with their employer, or financial reward. Accidental leaks mainly result from unintentional activities due to poor business process such as failure to apply appropriate preventative technologies and security policies, or employee oversight.

The purposes of data leak prevention and detection (DLPD) systems are to identify, monitor, and prevent unintentional or deliberate exposure of sensitive information in enterprise environment. Various technical approaches are used in DLPD targeting different causes of data leaks.[6] For example, several pioneering works[7,8] proposed to model normal database access behaviors in order to identify intruders and detect potential data breaches in relational databases. Basic security measures such as enforcing data use policies can safeguard sensitive information in storage. Traffic inspection is a commonly used approach to block sensitive data from being moved out of the local network.[9]

It is challenging for companies to protect data against information leakage in the era of big data. As data become one of the most critical components of an enterprise, managing and analyzing large amounts of data provides an enormous competitive advantage for corporations (e.g., business intelligence or personalized business service delivery). However, it also puts sensitive and valuable enterprise data at risk of loss or theft, posing significant security challenges to enterprises. The need to store, process, and analyze more and more data together with the high utilization of modern communication channels in enterprises result in an increase of possible data leakage vectors, including cloud file sharing, email, web pages, instant messaging, FTP (file transfer protocol), removable media/storage, database/file system vulnerability, camera, laptop theft, backup being lost or stolen, and social networks.

The purpose of this review paper is to highlight enterprise data leak threats, systematize solutions for data leak detection and prevention, and point out future research opportunities in this area. We first discuss categories of enterprise data leak threats, examine several data leak incidents in recent years, and summarize lessons learned from these incidents (*Enterprise Data Leak Threats* section). Then, we describe key DLPD techniques that have been developed in recent years and discuss the limitation of current DLPD approaches (*Data Leak Prevention and Detection Techniques* section). In particular, we highlight the challenges of DLPD systems in the era of big data and introduce a privacy preserving data leak detection system as a case study to address these challenges (*DLPD in the Big Data Era* section). Finally, we motivate future research in this area (*Further Research Opportunities* section).

ENTERPRISE DATA LEAK THREATS

The literature presents different taxonomies regarding data leak threats.[6,10] In this section, we use them to classify and describe major data leak threats. Then we review several enterprise data breach incidents and discuss lessons learned from these incidents.

Classification of Data Leak Threats

One approach to the classification of data leak threats is based on their causes, either intentionally or inadvertently leaking sensitive information. Another approach is based on which parties caused the leakage: insider or outsider threats. As shown in Figure 1, intentional leaks occur due to either external parties or malicious insiders. External data breaches are normally caused by hacker break-ins, malware, virus, and social engineering. For example, an adversary may exploit a system backdoor or misconfigured access controls to bypass a server's authentication mechanism and gain access to sensitive information. Social engineering (e.g., phishing) attacks become increasingly sophisticated against enterprises, by fooling employees and individuals into handing over valuable company data to cyber criminals. Internal data leakage can be caused by either deliberate actions (e.g., due to espionage for financial reward or employee grievances) or inadvertently mistakes (e.g., accidental data sharing by employees or transmitting confidential data without proper encryption). Hauer[11] proposed comprehensive criteria for characterizing totally 1259 data leakage incidents and analyzed data breaches reported in recent years. The results reveal that in over 60% of the data breaches were caused by insiders, highlighting that

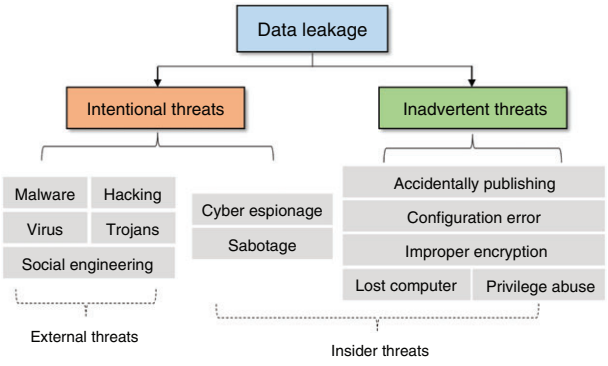

FIGURE 1 | Classification of enterprise data leak threats.

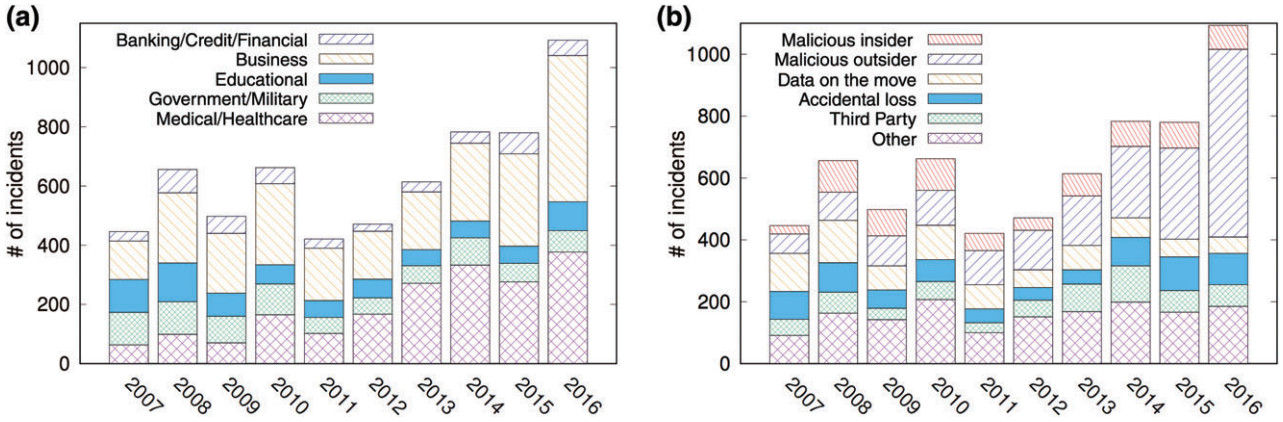

FIGURE 2 | Statistics of data leak incidents in recent years (Reprinted with permission from Ref 12. Copyright 2017 Identity Theft Resource Center). (a) Breaches by industry sector and (b) Breaches by type of occurrence

technological as well as nontechnological measures are both important in preventing data breaches.

Data leaks can also be characterized based on other attributes, such as by industry sector or by type of occurrence. As reported by Identity Theft Resource Center (ITRC)[12] in Figure 2, the total number of major data breach incidents (tracked by ITRC) keeps increasing in the past 5 years. For example, the number of data breach incidents in 2016 is 40% more than that in 2015. Figure 2(a) shows the stacked histogram plot of data breach incidents by industry sector. Business and medical/healthcare leaks take the majority of the leaks. In 2016, business data breach has 494 reports, taking 45.2% of the overall breaches, followed by medical/healthcare, representing 34.5% of the overall breaches with 377 incidents. Data breach by type of occurrence is illustrated in Figure 2(b), where the 'Other' category includes email/internet exposure or employee error and so forth. From the figure, the number of breaches caused by malicious outsider in 2016 takes around 55% of the overall incidents. Although different cybersecurity reports[5,13,14] may get different results due to using nonidentical datasets, all these reports, including ITRC's statistics, confirm the trend that insider threats emerge as the leading cause of enterprise data leak threats, with more than 40% of breaches perpetrated from inside a company.

Enterprise Data Leak Incidents

Over the past few years, massive enterprise data breaches have become a regular occurrence. Table 1 lists some notable data breaches in recent years, which shows that the consequences of an individual data breach could cause hundreds of millions of people having their personal information leaked, and incur financial loss of hundred million dollars. In the following, we describe several recent data breaches caused by external cyber attacks and insiders, respectively. In particular, we examine the Target data breach in detail,[17] which is a representative data leak incident as the result of outside attackers.

Internal Data Leak Incidents

There is an increase of accidental data leak incidents in recent years. For example, in October 2016, a staff

TABLE 1 | Massive Enterprise Data Leak Incidents in Recent Years (Data Source Is from the Dataset of World's Biggest Data Breaches[15])

Organization	Records	Breach Date	Type	Source	Industry	Estimated Cost
Anthem insurance	78 million	January 2015	Identify theft	Malicious outsider	Healthcare	$100 million
Yahoo	500 million	December 2014	Account access	State sponsored[1]	Business	$350 million
Home depot	109 million	September 2014	Financial access	Malicious outsider	Business	$28 million
JPMorgan chase	83 million	August 2014	Identify theft	Malicious outsider	Financial	$13 billion
Benesse	49 million	July 2014	Identify theft	Malicious insider	Education	$138 million
Korea credit bureau	104 million	January 2014	Identify theft	Malicious insider	Financial	$100 million
Target	110 million	November 2013	Financial access	Malicious outsider	Business	$252 million
Adobe System	152 Million	September 2013	Financial access	Malicious outsider	Business	$714 Million

[1] Announced by U.S. Department of Justice.[16]

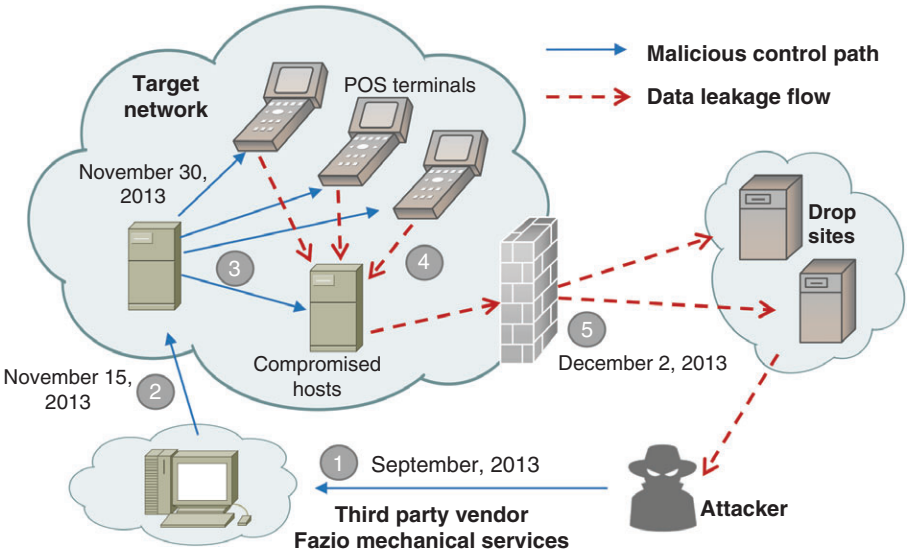

FIGURE 3 | Breakdown and analysis of the Target data breach.

from Australian Red Cross Blood Service accidentally placed the documents that contain more than 550,000 blood donors' personal information on an unsecured, public-facing directory of their website. The sensitive information relates to donors from 2010 to 2016, and includes names, addresses, and dates of birth as well as sexual activity, drug use, and medical histories. In 2011, A Texas State server accidentally published the personal information of 3.5 million citizens online for a year.

Various forms of malicious insider threat have been reported,[18] including extraction, duplication, exfiltration of sensitive data, eavesdropping, and packet sniffing, purposefully installing malicious softwares with backdoors. A high-impact data breach incident caused by insiders was the release of over 250,000 confidential documents of the United States diplomatic cables in 2010.[19] It was carried out by an internal entity using an external hard drive, and finally leaked to WikiLeaks. About 100,000 diplomatic cables were labeled 'confidential' and 15,000 cables had the higher classification 'secret.' This leak touched sensitive political issues and received a high level of attention from different governments over the world.[6] In 2013, an IT contractor for Vodafone Germany accessed to the telecom giant's database system and stole personal information and bank account details on as many as two million customers,[20] which may result in a possible increase in phishing attacks to customers. In response to the data breach, Vodafone reacted by changing the passwords and certificates of all administrators and completely reset the affected server to prevent further data leakage. As medical records across the nation become digitized, there have been many healthcare data leak incidents caused by insiders, which increased pressures to enhance cybersecurity practices for healthcare organizations. For example, in 2015, a former employee of UMass Memorial Medical Center was accused of having stolen up to 14,000 patient information such as names, date of births, addresses, and Social Security numbers from the hospital's billing application, which may have started 12 years ago.[21]

Detection of internal data leak incidents is extremely challenging, because internal breaches often involve users who have legitimate access to facilities and data. Their actions may not leave evidence due to their knowledge of organizations, possibly knowing how to bypass detection. With more and more covert channels and steganography tools available, malicious insiders make data breaches particularly difficult to detect. For example, malicious employees may bypass all enterprise security policies by concealing sensitive information to normal documents and sending them out via encrypted or covert channels. In the big data era, insiders are exposed to increasing amount of sensitive data, posing huge security challenges to organizations. To prevent unintentional or inadvertent data leakage, in addition to technological means, it is very important to increase user security awareness in workplace.[22]

External Data Leak Incidents

Many of the high-profile data breach incidents have resulted in organizations losing hundred millions of dollars. For examples, Yahoo and Target data

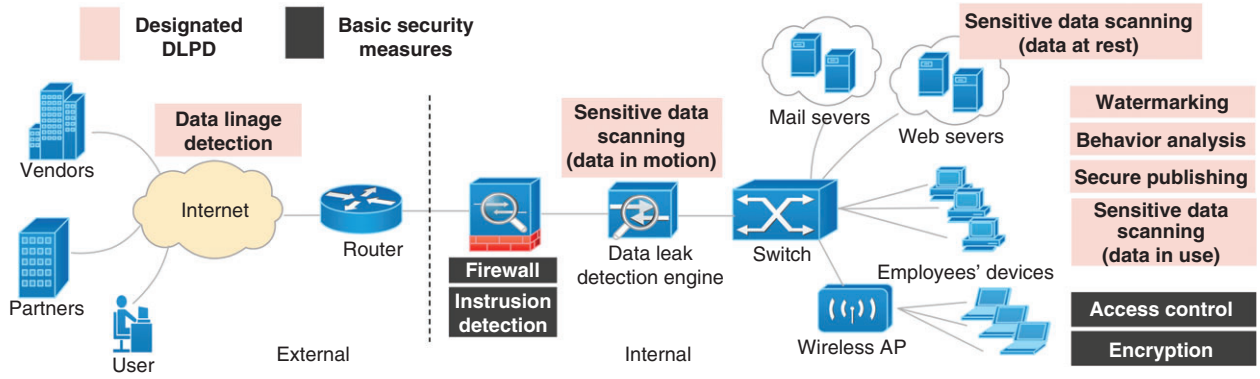

FIGURE 4 | There exist multiple points for deploying complementary data leak prevention and detection (DLPD) techniques in an enterprise environment.

breaches are among the biggest in history. Yahoo announced two huge data breaches in 2016. In the first incident, hackers compromised as many as 500 Million user accounts in late 2014. Later, in December 2016, Yahoo discovered another major cyber attack, more than 1 billion user accounts was compromised in August 2013, which is believed to be separate from the first one. After the data breaches, Verizon paid $350 million less than the originally planned sale price to acquire Yahoo. Between November 27 and December 18, 2013, cyber criminals breached the data security of Target Corporation, one of the nation's largest retailers. Later, it was announced by Target that personal information, including the names, addresses, phone numbers, email addresses, and financial information of up to 70 million customers, was stolen during the data breach.

Figure 3 illustrates how the Target data breach occurred. Attackers first compromised a third-party vendor Fazio Mechanical Services' system through a phishing attack in September 2013 (step 1). Fazio had access rights to Target's network for carrying out tasks like remotely monitoring energy consumption and temperatures of stores. In step 2, attackers penetrated into the Target networks, gaining access to vulnerable machines. Then, attackers compromised the POS (point of sales) systems and deployed the data stealing malware called BlackPOS on POS terminals (step 3), which could scan the memories of POS devices to read sensitive information. In step 4, stolen data were then encrypted and moved from POS devices to internal compromised hosts. Finally, attackers successfully moved data to drop sites out of the Target network (step 5).

Many external data leak threats like the Target data breach could potentially have been avoided, if adopting appropriate DLPD technical and administrative approaches. Target failed at detecting or preventing the breach at several points, and we identified four technical causes of the incident: (1) Target did not apply appropriate access control mechanisms on third-party partners, leading to the initial hacker break-ins. (2) It failed to segregate sensitive payment system from the rest of its networks. (3) Target did not harden the POS systems, allowing unauthorized software installation and configuration. (4) Although having firewalls and network intrusion prevention system (i.e., FireEye) in places, Target did not investigate into the security warnings issued by these security tools. A detailed analysis of the Target data breach can be found in Ref 17.

One positive take-home message for defenders from the Target case is that there are multiple places and opportunities to prevent or detect the data breach, e.g., anomaly-based traffic monitoring for recognizing abnormal destination and volume and access patterns, verification of code loading, restricting the access of nonssential business partners, and educating employees about phishing. Strategic deployment of proactive security defense mechanisms increases the difficulty level of attacks and reduces data leak risk.

DLPD TECHNIQUES

There exist a large number of DLPD techniques in the literature, where the majority of them are proposed by the research community and a small number of commercial products are coming from the industry. In this section, we review existing DLPD techniques and discuss their limitations.

Existing DLPD Techniques

We group existing DLPD techniques into two categories: basic security measures and designated DLPD

approaches. Unlike basic security mechanisms including firewall, antivirus software, intrusion detection, authentication, access control and encryption, DLPD systems are specially designated to deal with data leakage threats. The main task of DLPD is to identify, monitor, and protect confidential information from unauthorized access, which usually uses the actual content or surrounding context of the monitored data to detect potential leakage. In recent years, designated DLPD tools have gained popularity and will become an integral component of the enterprise' security suite.

Figure 4 illustrates typical techniques used to provide data leakage detection and prevention and their deployments in an enterprise system. Basic security measures such as secure data publishing, encrypting, and enforcing access rights to sensitive data safeguard data 'at rest,' which provides the first line of data leak mitigation. Firewalls limit access to the internal network. Intrusion detection systems (IDS) monitor computer and network activities to look for unauthorized intrusions. Antivirus may detect malware that steals confidential information before data is leaked, providing protection against internal attacks. IDS could be of help to detect malicious activities, but it normally suffers high false positive rates.[23] New mechanisms for protecting confidential files on a computer rely on the virtual machine technology.[24] Trusted computing technology is also used to provide a hardware-based root of trust for achieving content protection.[25]

We divide technological means employed in DLPD into two categories: *content-based analysis* and *context-based analysis*.

- Content-based (i.e., sensitive data scanning) approaches[9,26–30] inspect data content to protect unwanted information exposure in different states (i.e., at rest, in use, and in transit). Although content scanning can effectively protect against accidental data loss, it is likely to be bypassed by internal or external attackers such as by data obfuscation.
- Instead of trying to identify the presence of sensitive content, on the contrary, context-based approaches[7,8,31–35] mainly perform contextual analysis of the meta information associated with the monitored data or context surrounding the data. Some DLPD solutions are hybrid approaches that analyze both content and context.[36]

Since the main objective of DLPD is to identify content as sensitive, content-based methods normally achieve higher detection accuracy than pure context-based analysis, and thus the majority research efforts in this field focus on content analysis to detect sensitive data. As shown in Figure. 4, data scanning can be deployed at different points for protecting data in different stages. Scanning data at rest that are stored in servers enables enterprises to identify potential data leak risks within the internal organization. Monitoring data in use can avoid improper handling of sensitive data and prevent them from entering the enterprise network such as by blocking such traffic when detecting an attempt of transferring sensitive data. While monitoring network data streams in transit prevents confidential data from transmitting in and leaving the corporate network.

Content-Based Approaches

Content-based DLPD searches known sensitive information that resides on laptops, servers, cloud storage, or from outbound network traffic, which is largely dependent on data fingerprinting, lexical content analysis (e.g., rule-based and regular expressions), or statistical analysis of the monitored data. In data fingerprinting, signatures (or keywords) of known sensitive content are extracted and compared with content being monitored in order to detect data leaks, where signatures can either be digests or hash values of a set of data. Shapira et al.[27] proposed a fingerprinting method that extracts fingerprints from the core confidential content while ignoring nonrelevant (nonconfidential) parts of a document, to improve the robustness to the rephrasing of confidential content. Lexical analysis is used to find sensitive information that follows simple patterns. For example, regular expressions can be used for detecting structured data including social security numbers, credit card numbers, medical terms, and geographical information in documents.[37] Snort,[38] an open source network IDS, allows users to configure customized signatures and regular expression rules. Then sniffed packets in Snort will be compared against these signatures and rules to detect data leak attempts.

Statistical analysis mainly involves analyzing the frequency of shingles/n-grams, which are typically fixed-size sequences of contiguous bytes within a document. Another line of research includes the item weighting schemes and similarity measures in statistical analysis, where item weighting assigns different importance scores to items (i.e., n-grams), rather than treating them equally.

Collection intersection is a commonly used statistical analysis method in detecting the presence of sensitive data. Two collections of shingles are compared and the similarity score is computed between content sequences being monitored and sensitive data sequences that are not allowed to leave enterprise networks. For instance, the 3-gram shingles of a string abcdefgh include six elements { abc, bcd, cde,

def, efg, fgh}, where a sliding window is used in shingling the string. Given a content collection C_c and a sensitive data collection C_s, a detection algorithm computes the intersection rate $Irate \in [0, 1]$, which is defined as the sum of occurrence frequencies of all items appeared in the collection intersection $C_s \wedge C_c$ normalized by the size of $\min(|C_s| \wedge |C_c|)$. Figure 5 illustrates an example of calculating the similarity score of two 3-gram collections, where the sum of occurrence frequencies of items in $C_s \wedge C_c$ is 7, $\min(|C_s| \wedge |C_c|) = 10$, and thus the Irate is 0.7.

Recently, machine learning-based solutions have emerged to enable organizations to detect increasing amounts of confidential data that require protection. For example, Symantec utilizes vector machine learning (VML) technology in detecting sensitive information from unstructured data.[39] Through training, this approach can improve the accuracy and reliability of finding sensitive information continuously. Hart et al.[40] presented machine learning-based text classification algorithms to automatically distinguish sensitive or nonsensitive enterprise documents. Alneyadi et al.[41] used statistical analysis techniques to detect confidential data semantics in evolved data that appears fuzzy in nature or has other variations.

Context-Based Approaches

There have been a number of studies in profiling users' normal behaviors to identify intruders or insiders.[7,8,31–33,42,43] Instead of detecting the presence of sensitive data, Mathew et al.[33] proposed to model normal users' data access patterns and raise an alarm when a user deviates from the normal profile, in order to mitigate insider threat in database systems. Bertino et al.[7,8] proposed to detect anomalous access patterns in relational databases with a finer granularity based on mining database traces stored in log files. Their method is able to detect role intruders in database systems, where individuals holding a specific role behave differently from the normal behavior of the role. Senator et al.[43] presented a set of algorithms and methods to detect malicious insider activities, and demonstrated the feasibility of detecting the weak signals characteristic of insider threats on organizations' information systems. Costante et al.[31] addressed the problem of identifying and reacting to insider threats in data leak detection by monitoring user activities and detecting anomalous behavior. They presented a hybrid framework that combines signature-based and anomaly-based solutions. The anomaly-based component learns a model of normal user behavior to detect unknown and insider attacks, and then signature-based component automatically creates anomaly signatures (e.g., patterns of malicious activities) from alerts to prevent the execution of similar activities in the future. Gyrus[42] prevents malware from malicious activities such as manipulating a host machine to send sensitive data to outside parties, by capturing the semantics of user intent and ensuring that a system's behavior matches the user's intent. Maloof et al.[32] designed a system to monitor insider behavior and activity, in order to detect malicious insiders who operate within their privileges but engaging in activity that is outside the scope of their legitimate assignments.

Many of these context-based approaches are based on data mining or machine learning techniques. The advantage of machine learning-based approaches is that it does not need to precisely describe the anomalous activities by discovering outliers. Buczak et al.[44] surveyed popular machine learning and data mining techniques for cyber security anomaly detection. Both Buczak et al.[44] and Sommer et al.[45] highlighted that a challenge of applying machine learning or data mining for anomaly detection is the lack of training data.

Watermarking is used to prevent and detect data leaks, by marking data of interest entering and leaving a network. The presence of a watermark in an outbound document indicates potential data leak. It can also be used for forensics analysis (i.e., postmortem analysis) such as identifying the leaker after an incident.[34] In addition, trap-based defenses are also useful for the insider threat, which can entice and trick users into revealing their malicious intentions. For example,

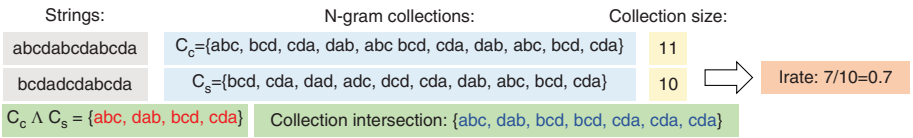

FIGURE 5 | An example of calculating the intersection rate between two 3-gram collections. Collection intersection counts duplicated items, whereas the set intersection does not.

Spitzner et al.[35] proposed to utilize honeypots for early detection of malicious insider threats. Their method implants honeytokens with perceived value in the network. Then, these honeytokens may direct the insider to more advanced honeypots, and discern whether the insider intention was malicious or not. Papadimitriou et al.[34] studied the problem of identifying guilty agents in the occurrence of data leakage. They proposed data allocation strategies for efficient assessing the likelihood that an agent is responsible for a leak. In addition, they consider the option of adding 'fake' content to the distributed data for identifying a leaker.

Limitations of Current DLPD Approaches

Different from traditional security measures, a practical DLPD system is desirable to meet the following requirements.[46] First, it only blocks data flows containing sensitive information, while accepting normal traffic in general. Second, it can defend against data loss by careless employees or malicious insiders. Third, even in the event that traditional security measures fail, it can prevent the malware or attacker from exfiltrating data from an organization's perimeter. Despite a plethora of research on DLPD, detecting and preventing enterprise data breaches remains an active research problem. Table 2 shows a summary of advantages and disadvantages of different DLPD techniques in academic research.

Signature-based detection is the most fundamental technique used in DLPD. In many instances, fingerprint databases are created by applying standard hash functions to documents that need protection. This approach is easy to implement and has a better coverage as it is able to detect the whole confidential content. However, data fingerprinting with conventional hashing can be easily bypassed and may yield false negatives when the sensitive data is altered or modified.[6] In addition, it may incur high computation cost when processing large content because it requires extensive data indexing and comparison between sensitive and normal data.

Many DLPD systems use regular expressions to perform exact and partial string matching. Regular expression-based comparison supports wildcards and thus can capture transformed data leaks to some extent. The problem with DLPSs using regular expressions analysis is that they offer limited data protection and yield high false positive rates. Thus, they are only suitable for detecting data leaks with predictable patterns.

For unstructured textual data, collection intersection is typically used to detect sensitive information. Since collection intersection preserves local features, it can tolerate a small amount of sensitive data modifications, e.g., inserted tags, character substitution, and lightly reformatted data. However, it suffers from high computation (i.e., time consuming) and storage cost. The basic n-gram-based detection may generate undesirable false alarms since the comparison is orderless. Shu et al.[30] proposed an alignment-based solution that measures the order of n-grams in collection intersection, which achieves more accurate detection rate than conventional string matching. To overcome the above issues, advanced content analysis such as machine learning-based methods have been proposed. Machine learning algorithms are also used in context-based DLPD approaches.[47] In the era of big data, the most severe problem of content-based DLPD approaches is scalability, i.e., they are not able to process massive content data in time.[26]

Behavior analysis for understanding user intention is important to mitigate the insider attack

TABLE 2 | Summary of Existing DLPD Techniques

Technique	Analysis	Pros	Cons
Fingerprinting[27,38]	Content	Simple, Better coverage	Very sensitive to data modification
Regular expressions[37,38]	Content	Simple, Tolerate certain noises	Limited data protection, High false positive
Collection intersection[9,26,30]	Content	Wide data protection, Capture local features	High computation and storage cost, Inapplicable to evolved or obfuscated data
Machine learning[39,40,47]	Content/ Context	Resilient to data modifications, High accuracy	Large training data, Complicated
Behavior analysis[7,8,31–33]	Context	Mitigate insider threats	Large training data, High false positives
Watermarking[34]	Context	Forensics analysis	Vulnerable to malicious removal or distortion
Honeypots[35]	Context	Detect malicious insiders	Limited applications

DLPD, data leak prevention and detection.

problem. Insider threat detection has attracted significant attention in recent years. A plethora of behavior models as well as audit sources are available in the literature.[48] However, existing behavior analysis-based approaches are prone to errors because of the temporal dynamics of context information, and thus leading to high false positive and low detection rates. Watermarking is vulnerable to malicious removal or distortion and may involve modification of the original data, which limit its practical application in DLPD. Honeypots approach has its inherent drawback that the insider may not ever use or interact with the honeypots.

Although existing DLPD techniques are effective at preventing accidental and plain-text leaks, they are fundamentally unable to identify encrypted or obfuscated information leaks. Borders et al.[49] addressed the threat of a hacker or malicious insider steal data without being detected by hiding it in the noise of normal outbound Web traffic. The authors presented an approach for quantifying information leak capacity in network traffic by discounting fields that are repeated or constrained by the protocol, making it possible to identify stealthy information leaks. TM-Score[50] and M-Score[51] assign sensitivity scores to tabular data and textual data, respectively. Using these two measures, organizations are able to predict the ability of an insider to maliciously exploit the exposed data, estimate the risk of data exposure, and further mitigate leakage or misuse incidents of sensitive data.

Gugelmann et al.[52] conducted a systematic performance evaluation of three state-of-the-art DLPD tools to examine whether they can effectively prevent data leakage in Web traffic. Results demonstrate that these solutions of major vendors can prevent accidental data leakage, while insiders or malware can easily bypass the protection measures such as by obfuscating data. Despite many industrial solutions in DLPD, they are insufficient to protect against malicious data breaches. Although context analysis has the potential to solve this problem, building effective insider detection systems remains an open challenge.

DLPD IN THE BIG DATA ERA

Alneyadi et al.[6] summarized several challenges faced by DLPD systems in their survey paper, including the increase in leaking channels, human factors, encryption, and steganography, etc. In this section, we highlight the challenges in data leakage detection systems in the era of big data. We also present a privacy preserving data leak detection system as a case study to address these challenges.

Challenges

While the rise of big data yields tremendous opportunities for enterprises, data leak risk inevitably arises because of the ever-growing data volumes within corporate systems. For the same reason, data breach incidents will become more damaging to enterprises. In many cases, sensitive data are shared among various stakeholders, e.g., business partners and customers. Cloud file sharing and external collaboration with companies, which are becoming more common for today's enterprises, make the data leakage issue even worse. On the other hand, as workforce is becoming mobile, employees working from outside the organization's premises raise the potential for data leaks. In addition, in big data environments, motivations behind cyberattacks on stealing confidential enterprise data are dramatically increased with bigger payoffs and more recognition from a single attack. These factors pose a greater challenge of detecting unauthorized use, access, and disclosure of confidential enterprise data. Here, we list several technical challenges for data leak detection in the era of big data.

- Scalability: the ability to process large content, e.g., megabytes to terabytes, and can be deployed in distributed environments where the operating nodes are owned by third-party service providers, e.g., data leak detection in the Cloud. Scalability is the key to efficiently processing massive enterprise-scale amounts of data. A scalable solution can also reduce the data processing delay and achieve early data leak detection.
- Privacy Preservation: the ability to preserve the confidentiality of sensitive data from the DLPD provider or any attacker breaking into the detection system. Privacy is a major concern when outsourcing data leak detection to third-party vendors.
- Accuracy: achieving low false negative/positive rates for the detection. The distributed nature of big data environments poses a challenge in accurate leak detection. The outsourced data to third-party vendors may be transformed or modified by different users or applications, e.g., insertions of metadata or formatting tags, substitutions of characters for formatting purposes. As a result, it reduces the accuracy of content-based approaches.
- Timeliness: immediately detect and respond to data breaches before they cause damages. The volume, variety, and velocity of big data bring both opportunities and challenges for nearly real-time identifying data leak threats.

A Case Study

To address the above challenges, we now introduce a privacy preserving data leak detection system as a case study, named MapReduce-based Data Leak Detection (MR-DLD).[26] It utilizes the MapReduce[53] distributed computing framework to inspect sensitive content for inadvertent data leak detection, and can be deployed either in local computer clusters or in the Cloud. MapReduce has been widely used for distributed data intensive applications such as spam filtering or traffic/log analysis. MR-DLD implements the collection intersection algorithm (introduced in *Content-Based Approaches* section) based on MapReduce framework for detecting the occurrences of sensitive data patterns in massive-scale content in data storage or network transmission. The design goal of MR-DLD is to offer DLPD provider solutions to scan massive content for sensitive data exposure and minimize the possibility that the DLPD provider derives sensitive information during the data scanning.

Due to limited computation and storage capability, data owners (e.g., organizations) may outsource data leak detection tasks to third parties. However, the DLPD provider might be honest-but-curious (*aka.* semi-honest), which means it follows the prescribed protocol but may attempt to reconstruct the sensitive data. To prevent exposure of sensitive data during content scanning, it is preferable that data owners perform data transformations before releasing to MapReduce nodes, rather than sending plaintext content to third parties. In MR-DLD, the collection intersection is computed based on Rabin fingerprints,[54] i.e., one-way hash values of n-grams, to provide strong-yet-efficient confidentiality protection for the sensitive content.

Figure 6 illustrates the high-level workload distribution between the data owner and DLPD provider. When DLPD provider receives both the sensitive fingerprint collections and content fingerprint collections. It deploys MapRecuce framework and compares the content collections with the sensitive collections using two-phase MapReduce algorithms (i.e., map and reduce operations, respectively). By computing the intersection rate of each content and sensitive collections pair, it outputs whether the sensitive data was leaked and reports all the data leak alerts to data owner, given a predefined threshold. Data owner receives alerts with a set of tuples content and sensitive collection pairs, retransforms them to suspicious content segments and the plaintext sensitive sequences tuples, and finally identifies the leak occurrences. Details of MapReduce algorithms are not described here, and interested readers are referred to the paper.[26] Interested readers are also referred to another relevant work that used GPU to accelerate screening of transformed data leaks.[30]

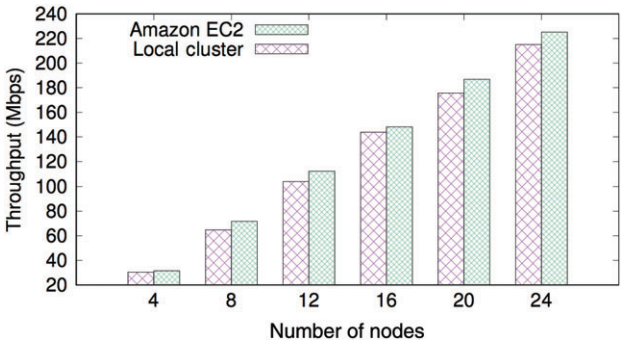

FIGURE 7 | MR-DLD throughput on a local cluster and Amazon EC2.

For scalability evaluation, 37 GB content was processed with different numbers of nodes, ranging from 4 from 24. The experiments were deployed both on the local cluster and on Amazon EC2. Figure 7 demonstrates that the system scales well as the throughput increases with the number of nodes. The peak throughput observed is 215 Mbps on the local cluster and 225 Mbps on Amazon EC2, and EC2 cluster yields 3–11% performance improvement. This improvement is partly due to the larger memory on Amazon EC2 platform.

FURTHER RESEARCH OPPORTUNITIES

As discussed early, there are still many research issues and opportunities where further research efforts are required, especially as the enterprise data

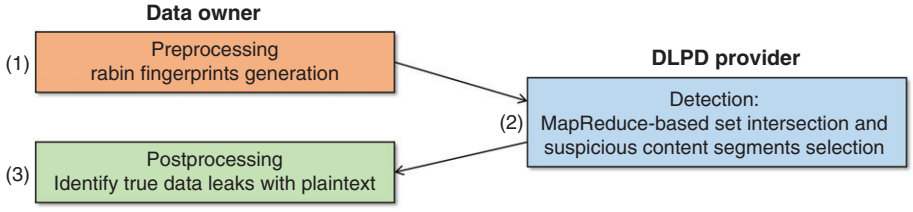

FIGURE 6 | Workload distribution between the data owner and data leak prevention and detection (DLPD) provider.

volumes are rapidly increasing. We highlight few of them in this section.

- Deep Learning for Insider Threat Detection: In big data settings, where a large volume of data from heterogeneous sources are generated, data mining and machine learning techniques will be increasingly used in DLPD.[55] Deep learning techniques such as the Deep Neural Network have been used to detect anomalies in different applications.[56] Such techniques can be applied to both content and context analyses in DLPD, which will be able to not only reveal stealthy data leaks, but also improve accuracy and achieve timely protection. Deep learning may also help close the semantic gap often encountered in the insider threat detection. The semantic gap is between the high-level user intentions and the low-level machine events. User intentions are the most pertinent to detecting insiders, however, they are not directly measurable. On the contrary, machine events are directly observable, unfortunately, they are not meaningful and need to be mapped to corresponding user intentions. Similar semantic gaps exist in many other research problems, e.g., capturing image semantics based on pixels.[57–59] Deep learning techniques have recently shown promises in solving complex sequence-to-sequence translation problems in natural languages.[60] Training deep learners to infer sequences of user intentions based on sequences of machine events is an extremely interesting direction.
- DLPD as a Cloud Service: The advent of cloud computing offers a new option for conducting data leak detection. Enterprises may outsource their data processing to third-party service providers, which brings about data privacy concerns. Collection intersection approach is based on the similarity of two sets with their element frequency information. Therefore, it might be vulnerable to frequency analysis if the sensitive data is outsourced to a third party and the third party has enough background frequency information of the n-gram. Privacy preserving data leak detection algorithms are needed to resist strong attacks. An important research direction is for the cloud service provider is to achieve scalability without reducing detection accuracy and incurring significant delay when processing large-scale datasets. Spark[61] is able to process streaming dataset by trucking data streams into small data segments. It is compatible with the MapReduce-based detection approach. However, the small content segments may miss the real leaks if the leak happens across multiple data segments, where increasing the size of data segments also increases transmission delay. Flink[62] is another stream data processing platform that may be used to build enterprise-scale data leak detection.
- Monitoring Encrypted Channels: Most existing DLPD approaches discussed are vulnerable to large alteration of the original data, and thus are inapplicable to evolved, obfuscated, or encrypted data. Encrypted traffic inevitably renders existing content-based detection useless. While deploying monitors outside, the encrypted channel can partially mitigate the problem, future DLPD solution needs a way to monitor encrypted channels in order to effectively detect stealthy data leaks. A possible direction is the use of data flow tracking[63] or differential analysis. For example, researchers recently leveraged differential analysis technique to achieve obfuscation resilient privacy leak detection on smartphone platforms.[64] String matching on encrypted data[65] has been one of the hot research areas in the last decade. Techniques in this area may also be used in future DLPD to detect the transfer of sensitive information on encrypted channels.
- Benchmarks for DLPD: Sommer et al.[45] pointed out that the lack of training data is one of the challenges for applying machine learning to network anomaly detection, which also applies to DLPD. As machine learning techniques are being increasingly used, academic research in DLPD lacks common datasets for testing and evaluation, making it hard to compare with the state-of-the-art solutions and perform sound evaluation. The research community needs to provide mechanisms to incentivize data sharing and benchmark preparation effort.

CONCLUSION

Preventing and detecting data leaks require constant effort and investment from organizations. In this paper, we have presented a review of data leak threats and key techniques for DLPD. We described the state-of-the-art DLPD techniques that have been developed in recent years. While existing survey papers[6,10] provide more thorough descriptions of these techniques, in this review article, we highlighted the challenges that still need to be addressed,

particularly in the big data era. We also pointed out several promising research directions for reducing data breach risks in enterprise environments. We found that the directions of data leak detection as a cloud service and deep learning-based anomaly detection for insider threat are particularly promising.

REFERENCES

1. 2016 cost of data breach study: global analysis. 2017. Available at: https://www-03.ibm.com/security/data-breach. (Accessed March 1, 2017).
2. Cybercrime will cost businesses over $2 trillion by 2019. 2017. Available at: https://www.juniperresearch.com/press/press-releases/cybercrime-cost-businesses-over-2trillion. (Accessed March 1, 2017).
3. The target and other financial data breaches: frequently asked questions. 2017. Available at: https://fas.org/sgp/crs/misc/R43496.pdf. (Accessed March 1, 2017).
4. Yahoo says 500 million accounts stolen. 2017. Available at: http://money.cnn.com/2016/09/22/technology/yahoo-data-breach/. (Accessed March 1, 2017).
5. Data exfiltration study: Actors, tactics, and detection. 2017. Available at: https://www.mcafee.com/us/resources/reports/rp-data-exfiltration.pdf. (Accessed March 1, 2017).
6. Alneyadi S, Sithirasenan E, Muthukkumarasamy V. A survey on data leakage prevention systems. *J Netw Comput Appl* 2016, 62(C):137–152.
7. Bertino E, Terzi E, Kamra A, Vakali A. Intrusion detection in RBAC-administered databases. In: *21st Annual Computer Security Applications Conference (ACSAC'05)*, 2005, 1–10.
8. Kamra A, Terzi E, Bertino E. Detecting anomalous access patterns in relational databases. *VLDB J* 2008, 17:1063–1077.
9. Shu X, Yao D, Bertino E. Privacy-preserving detection of sensitive data exposure. *IEEE Trans Inform Forensic Secur* 2015, 10:1092–1103.
10. Shabtai A, Elovici Y, Rokach L. *A Survey of Data Leakage Detection and Prevention Solutions*. Berlin, Heidelberg: Springer Science & Business Media; 2012.
11. Hauer B. Data and information leakage prevention within the scope of information security. *IEEE Access* 2015, 3:2554–2565.
12. Identity Theft Resource Center. 2017. Available at: http://www.idtheftcenter.org/. (Accessed March 1, 2017).
13. Data breach investigations report. 2017. Available at: http://www.verizonenterprise.com/verizon-insights-lab/dbir/. (Accessed March 1, 2017).
14. Bertino E. Security threats: protecting the new cyber-frontier. *Computer* 2016, 49:11–14.
15. World's biggest data breaches. 2017. Available at: http://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/. (Accessed March 1, 2017).
16. DOJ: 2 Russian spies indicted in yahoo hack. 2017. Available at: http://www.cnn.com/2017/03/14/politics/justice-yahoo-hack-russia/index.html. (Accessed March 1, 2017).
17. Shu X, Tian K, Ciambrone A, Yao D. Breaking the target: an analysis of target data breach and lessons learned. ArXiv e-prints. 2017.
18. Phyo AH, Furnell SM. A detection-oriented classification of insider IT misuse. *Computers & Security* 2002: 21:62–73.
19. Data breach investigations report. 2017. Available at: https://en.wikipedia.org/wiki/United_States_diplomatic_cables_leak. (Accessed March 1, 2017).
20. Insider steals data of 2 million Vodafone Germany customers. 2017. Available at: http://www.securityweek.com/attacker-steals-data-2-million-vodafone-germany-customers. (Accessed March 1, 2017).
21. UMass memorial insider breach went on for 12 years. 2017. Available at: https://www.observeit.com/blog/umass-memorial-insider-breach-went-12-years. (Accessed March 1, 2017).
22. Colwill C. Human factors in information security: the insider threat - who can you trust these days? *Inf Secur Tech Rep* 2009, 14:186–196.
23. Julisch K, Dacier M. Mining intrusion detection alarms for actionable knowledge. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, July 23–26. New York, NY: ACM; 2002, 366–375.
24. Borders K, Weele EV, Lau B, Prakash, A. Protecting confidential data on personal computers with storage capsules. In: *Proceedings of the 18th Conference on USENIX Security Symposium*, Montreal, Canada. Berkeley, CA: USENIX Association; 2009, 367–382.
25. Alawneh M, Abbadi IM. Preventing information leakage between collaborating organisations. In: *Proceedings of the 10th International Conference on Electronic Commerce*, ICEC '08, Innsbruck, Austria. New York, NY: ACM; 2008, 38:1–38:10.
26. Liu F, Shu X, Yao D, Butt AR. Privacy-preserving scanning of big content for sensitive data exposure with MapReduce. In: *Proceedings of the 5th ACM Conference on Data and Application Security and*

Privacy, CODASPY 2015, San Antonio, TX, 2–4 March, 2015, 195–206.

27. Shapira Y, Shapira B, Shabtai A. Content-based data leakage detection using extended fingerprinting. *CoRR* abs/1302.2028. 2013.
28. Shu X, Yao D. Data leak detection as a service. In: *Proceedings of the 8th International Conference on Security and Privacy in Communication Networks (SecureComm)*, Padua, Italy, September, 2012, 222–240.
29. Shu X, Zhang J, Yao D, Feng W. Rapid and parallel content screening for detecting transformed data exposure. In: *Proceedings of the Third International Workshop on Security and Privacy in Big Data (BigSecurity)*, Hongkong, China, April, 2015.
30. Shu X, Zhang J, Yao DD, Feng WC. Fast detection of transformed data leaks. *IEEE Trans Inform Forensic Secur* 2016, 11:528–542.
31. Costante E, Fauri D, Etalle S, Hartog JD, Zannone N. A hybrid framework for data loss prevention and detection. In: *2016 I.E. Security and Privacy Workshops (SPW)*, San Jose, USA, 2016, 324–333.
32. Maloof MA, Stephens GD. ELICIT: A System for Detecting Insiders Who Violate Need-to-know In: *Proceedings of the 10th International Conference on Recent Advances in Intrusion Detection*. Gold Goast, Australia. Berlin, Heidelberg: Springer-Verlag; 2007, 146–166.
33. Mathew S, Petropoulos M, Ngo HQ, Upadhyaya S. A data-centric approach to insider attack detection in database systems. In: *Proceedings of the 13th International Conference on Recent Advances in Intrusion Detection*, RAID'10, Ottawa, Ontario, Canada. Berlin, Heidelberg: Springer-Verlag; 2010, 382–401.
34. Papadimitriou P, Garcia-Molina H. Data leakage detection. *IEEE Trans Knowl Data Eng* 2011, 23:51–63.
35. Spitzner L. Honeypots: catching the insider threat. In: *19th Annual Computer Security Applications Conference (ACSAC)*, Las Vegas, NV, USA. Washington, DC: IEEE Computer Society; 2003, 170–179.
36. Carvalho VR, Cohen WW. Preventing Information Leaks in Email. In: *Proceedings of the 2007 SIAM International Conference on Data Mining*, Minneapolis, USA. Philadelphia: SIAM; 2007, 68–77.
37. Regular expressions openDLP. 2017. Available at: http://code.google.com/p/opendlp/wiki/RegularExpressions. (Accessed March 1, 2017).
38. Roesch M. Snort—lightweight intrusion detection for networks. In: *Proceedings of the 13th USENIX Conference on System Administration*, LISA '99, Seattle, Washington. Berkeley, CA: USENIX Association; 1999, 229–238.
39. Machine learning sets new standard for data loss prevention: Describe. 2017. Available at: http://eval.symantec.com/mktginfo/enterprise/white_papers/b-dlp_machine_learning.WP_en-us.pdf. (Accessed March 1, 2017).
40. Hart M, Manadhata P, Johnson R. Text classification for data loss prevention. In: *Proceedings of the 11th International Conference on Privacy Enhancing Technologies*, PETS'11, Waterloo, ON, Canada. Berlin, Heidelberg: Springer; 2011, 18–37.
41. Alneyadi S, Sithirasenan E, Muthukkumarasamy V. Detecting data semantic: a data leakage prevention approach. In: *IEEE Trustcom*. New York: IEEE; 2015, 910–917.
42. Jang Y, Chung SP, Payne BD, Lee W. Gyrus: a framework for user-intent monitoring of text-based networked applications. In: *21st Annual Network and Distributed System Security Symposium (NDSS)*, San Diego, California, USA. Reston, VA: Internet Society; 2014.
43. Senator TE, Goldberg HG, Memory A, Young WT, Rees B, Pierce R, Huang D, Reardon M, Bader DA, Chow E, et al. Detecting insider threats in a real corporate database of computer usage activity. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Chicago, Illinois, USA. New York, NY: ACM; 2013, 1393–1401.
44. Buczak AL, Guven E. A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Commun Surv Tutor* 2016, 18:1153–1176.
45. Sommer R, Paxson V. Outside the closed world: on using machine learning for network intrusion detection. In: *Proceedings of the 2010 I.E. Symposium on Security and Privacy*, Oakland, California, USA. Washington, DC: IEEE Computer Society; 2010, 305–316.
46. Gugelmann D. On data and privacy leakage in web traffic. PhD thesis, *ETH-Zürich*, 2015.
47. Brdiczka O, Liu J, Price B, Shen J, Patil A, Chow R, Bart E, Ducheneaut N. Proactive insider threat detection through graph learning and psychological context. In: *Proceedings of the 2012 IEEE Symposium on Security and Privacy Workshops*, Oakland, California, USA. Washington, DC: IEEE Computer Society; 2012, 142–149.
48. Salem MB, Hershkop S, Stolfo SJ. A Survey of Insider Attack Detection Research. In: Insider Attack and Cyber Security: Beyond the Hacker. Boston, MA: Springer US; 2008, 69–90.
49. Borders K, Prakash A. Quantifying information leaks in outbound web traffic. In: *Proceedings of the 2009 30th IEEE Symposium on Security and Privacy* Oakland, California, USA. Washington, DC: IEEE Computer Society; 2009; 129–140.

50. Harel A, Shabtai A, Rokach L, Elovici Y. M-score: a misuseability weight measure. *IEEE Trans Dependable Secure Comput* 2012, 9:414–428.
51. Vartanian A, Shabtai A. Tm-score: a misuseability weight measure for textual content. *IEEE Trans Inform Forensic Secur* 2014, 9:2205–2219.
52. Gugelmann D, Studerus P, Lenders V, Ager B. Can content-based data loss prevention solutions prevent data leakage in web traffic? *IEEE Secur Priv* 2015, 13:52–59.
53. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. *Commun ACM* 2008, 51:107–113.
54. Broder AZ. Some applications of Rabin's fingerprinting method. In: *Sequences II: Methods in Communications, Security, and Computer Science*. Berlin, Heidelberg: Springer; 1993, 143–152.
55. Sommer R, Paxson V. Outside the closed world: on using machine learning for network intrusion detection. In: *Proceedings of the 2010 IEEE Symposium on Security and Privacy*, Oakland, California, USA. Washington, DC: IEEE Computer Society; 2010, 305–316.
56. Schmidhuber J. Deep learning in neural networks: an overview. *Neural Netw* 2015, 61:85–117.
57. Lin L, Ravitz G, Shyu ML, Chen SC. Correlation-based video semantic concept detection using multiple correspondence analysis. In: *Proceedings of the 2008 Tenth IEEE International Symposium on Multimedia*, Berkeley, CA, USA. Washington, DC: IEEE Computer Society; 2008, 316–321.
58. Mojsilovic A, Rogowitz B. Capturing image semantics with low-level descriptors. In: *Proceedings 2001 International Conference on Image Processing, Thessaloniki, Greece*. 2001, 1, 18–21.
59. Sadanand S, Corso JJ. Action bank: a high-level representation of activity in video. In: *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Rhode Island, USA*. Washington, DC: IEEE Computer Society; 2012, 1234–1241.
60. Johnson M, Schuster M, Le QV, Krikun M, Wu Y, Chen Z, Thorat N, Viégas FB, Wattenberg M, Corrado G, et al. Google's multilingual neural machine translation system: enabling zero-shot translation. *CoRR abs/1611.04558*, 2016.
61. Apache spark: lightning-fast cluster computing. 2017. Available at: http://spark.apache.org/. (Accessed March 1, 2017).
62. Apache Flink: scalable stream and batch data processing. 2017. Available at: https://flink.apache.org/. (Accessed March 1, 2017).
63. Priebe C, Muthukumaran D, O' Keeffe, D, Eyers D, Shand B, Kapitza R, Pietzuch P. Cloudsafetynet: detecting data leakage between cloud tenants. In: *Proceedings of the 6th Edition of the ACM Workshop on Cloud Computing Security*, CCSW '14, Scottsdale, Arizona, USA. New York, NY: ACM; 2014, 117–128.
64. Continella A, Fratantonio Y, Lindorfer M, Puccetti A, Zand A, Kruegel C, Vigna G. Obfuscation-resilient privacy leak detection for mobile apps through differential analysis. In: *Proceedings of the ISOC Network and Distributed System Security Symposium (NDSS)*, San Diego, California, USA. Reston, VA: Internet Society; 2017, 1–16.
65. Kuzu M, Islam MS, Kantarcioglu M. Efficient similarity search over encrypted data. In: *Proceedings of the 2012 IEEE 28th International Conference on Data Engineering*, ICDE '12, *Arlington, Virginia, USA*. Washington, DC: IEEE Computer Society; 2012, 1156–1167.