DOCTORAL CONSORTIUM

On Data Leakage Prevention And Machine Learning

Jan Domnik[1] & Alexander Holland[2]

[1] Universidad Católica San Antonio de Murcia (UCAM), Spain.
E-mail: ucam@domnik.es
[2] FOM University of Applied Sciences, Essen, Germany.
E-mail: alexander.holland@fom.de

Abstract An analyst in the field of Data Leakage Prevention (DLP) usually inspects suspicious file transfers which are called events. First of all, the data in question is classified. Then, the context of the transfer is determined. After this, the analyst decides whether the transfer was legitimate or not. This process is widely known as triage. It is monotonous, costly and resource-intensive. Therefore the following question arises; could modern DLP-Software utilize machine learning algorithms in order to automate the triage process? Further, this begs the question, which structural and organisational processes are necessary inside an organisation to automate that process. In this case, it could significantly enhance the quality of DLP practices and take work from the much needed human resources in the field of IT security. Further, DLP systems (today usually used in bigger organisations) could become more attractive and more specifically affordable for small- and medium-sized organisations.

Keywords: DLP, data leakage prevention, data loss prevention, machine learning, IT-Security.

University of Maribor Press

DOI https://doi.org/10.18690/um.fov.4.2022.45
ISBN 978-961-286-616-7

1 Introduction

The old saying that data is the oil of the 21st century is not true of most data-sets upon closer inspection. Nevertheless, most organisations create, process or save highly sensitive data that should not be leaked by an internal or external attacker.

Ensuring the confidentiality, integrity and availability of sensitive data is the basis of IT security, which is included in the much bigger field of information security, which in turn, ensures the protection of an organisation's values and assets.

By taking a closer look at the security objective confidentiality, it quickly becomes clear that sensitive data is not exclusively exposed to risks outside of an organisation. Negligent acts and internal attacks can also lead to a data loss incident. To detect and prevent unauthorized data transfers from an employee (by accidence) or an internal attacker, Data Leakage Prevention (DLP) systems are in place.

2 Data Confidentiality

2.1 Data Classification

To protect the confidentiality of sensitive data effectively, most well-known international information security standards (e.g. ISO/IEC 27000) recommend the classification of all data that is processed by an organisation. Based on common best practices, organisations establish policies that normally differentiate between public, internal, restricted and highly-restricted data.

Further policies are typically required for handling each individual data type outlined in the previous paragraph. Those policies also commonly differentiate between Data at Rest and Data in Motion. The definition of Data at Rest (e.g. (Broadcom, 2022)) includes all data that is permanently stored, e.g. files that were stored on hard-disks or network drives. Data in Motion describes data which is in transit e.g. web-uploads, e-mails, print-jobs or file-transfers via USB.

To provide effective data leakage prevention, the described policies for data classification and data handling need to be in place. In short, they provide the necessary controls that are technically enforced by data leakage prevention.

2.2 Data Leakage Prevention

In this paper, Data Leakage Prevention is defined as the detection of intentional and unintentional violations against policies regarding Data in Motion and Data at Rest. Therefore DLP can be understood as a common task of an organization's Security Operations Center (SOC); where outgoing data is analysed against a specified set of rules.

Using complex hard- and software solutions, which are provided by a small variety of software manufacturers, is the most popular approach of detecting policy violations. Although each solution has its own strengths and weaknesses, they all provide some kind of common feature set. Furthermore they all trigger an event for each potential policy violation that is detected by the various mechanisms(AlKilani et al., 2019; Alneyadi et al., 2016; Gugelmann et al., 2015; Ouellet and McMillan, 2011; Ullah et al., 2018).

DLP Software is commonly based on one of the two following concepts.

On the one hand there is the content-based approach of detecting policy violations. It scans transferred or stored data using (partially highly sophisticated) patterns which are called rulesets. For example, an event could be triggered by specific keywords, file-types, file-sizes or renamed file-endings. A smart black- and whitelisting of the events may significantly reduce the amount of false positive events. Several or all events are determined by an analyst later. But this is also the bottleneck of the content-based approach: a lot of false positive events are generated.

To separate the wheat from the chaff, each event gets a risk-score which is calculated by a heuristic that tries to apply an appropriate risk-level. But implementing a meaningful risk-score is difficult if not impossible in practice. Because of this, the content-based approach often still results in a huge workload for an analyst. It is not unusual, that - depending on the organisation's risk appetite - the false positive rate of all the collected events far exceeds 90 percent. However, the false negative rate still remains unknown. All of this makes the content-based approach costly and resource-intensive.

An alternative approach, namely the behaviour-based approach, is often proposed to detect policy violations. Using correlations in a user's behaviour, the algorithm tries to detect anomalies which might contain a policy violation. This approach only examines the user's behaviour, while, for example, leaving suspicious e-mail-attachments out of the scope of anomalous behaviour. The disadvantages of this approach seem to be obvious: the idea of a correlation based analysis is in practice undermined by the fact that this approach tries to use correlations to detect causal links. For example, if an employee typically sends a set number of personal e-mails to a personal e-mail address each week, and suddenly this pattern changes (or not), is this truly indicative (or not) of a policy violation?

Of course, articles describing this approach(Faiz et al., 2020), where an e-mail attachment is just seen as a yes or no flag, have been shown to have some scientific backing. On top of this, well defined and barely changing processes can be protected against many cases of data leakage using this method, so it is predestined to be used in high-security environment. On the other hand, the number of processes required and the complexity of these processes is often unspecified (especially in small- to medium-sized companies), making it impossible to follow this kind of approach. So it is not surprising that the behaviour-based approach is often seen as a kind of snake oil in these environments.

Both approaches have the following in common: the decision about escalating an event as an incident to the management level or determining the final classification of an event must be undertaken by a human being. In order to be able to make an informed decision, the data in question has to be analysed, classified and set into the correct context. This process is known as triage.

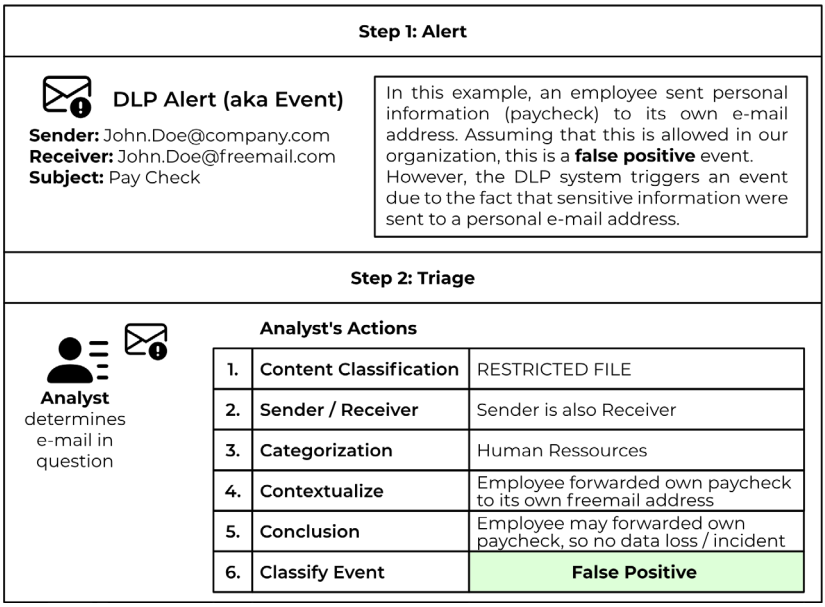

Figure 1: An example on how the triage process is performed

3 Thresholds And False Positive Events

As described, operational costs of a DLP system are not only expensive because of licensing, but also because of the resources needed for its administration. Further, each event that has been assigned a specified risk must be triaged by a human analyst.

If a company decides that policy violations should not be determined by a non-transparent algorithm, the content-based approach is often chosen for DLP activities. In this case, thresholds for the DLP rule sets have to be defined. If these rules are really strict (e.g. each E-Mail above 3 KB in size sent to a freemail account, triggers an event), lots of false positives are created. If they are less strict, the risk of an undetected DLP incident rises.

The resources needed to perform the triage are linearly increasing with the amount of triggered events. Or, in other words: The greater the DLP risk an organisation is willing to accept, the less the company will have to pay.

As noted earlier, only a few organisations have structures and processes in place that allow the automated detection of DLP incidents using heuristics. However, an automated incident detection system (or a reliable risk rating of events) can reduce DLP risks of an organisation, however, it may increase the time spent by an analyst reviewing false positive events. On top of this, DLP systems could be made more attractive for small- and medium-sized organisations which don't actually have the software, hardware and structures in place to perform a cost efficient risk treatment on DLP risks.

4 Is an automation of the triage process possible?

Regarding several data sources that could be used to apply a machine learning algorithm (e.g. organisation chart, active directory logs, e-mail history, employee address books, proxy logs, etc.), the research gap that is actually addressed is the following:

How far - and under which circumstances - can the DLP triage be automized by using appropriate machine learning algorithms?

So, in detail, the following topics are actually covered:

- Which data leakage scenarios can be detected automatically and which data sources need to be available for that?
- If not possible: How far can well established machine learning algorithms support the automation approaches?
- Which knowledge transfer methods can be used to continuously improve a DLP system as well as the surrounding processes?

5 Approach

1. **Get / Explore / Prepare a test set**. If there is no suitable DLP test set available, an training set needs to be created (e.g. from (CALO Project, 2011))
2. **Include data leakage attempts:** Put e-mails into the data set which include »normal« and advanced data leakage attempts, based on a threat model
3. **Use machine learning to perform the triage:** Retrace state of the art papers to create suitable ways to automate the triage

4. **Compare:** ...with the classic triage as it is performed nowadays

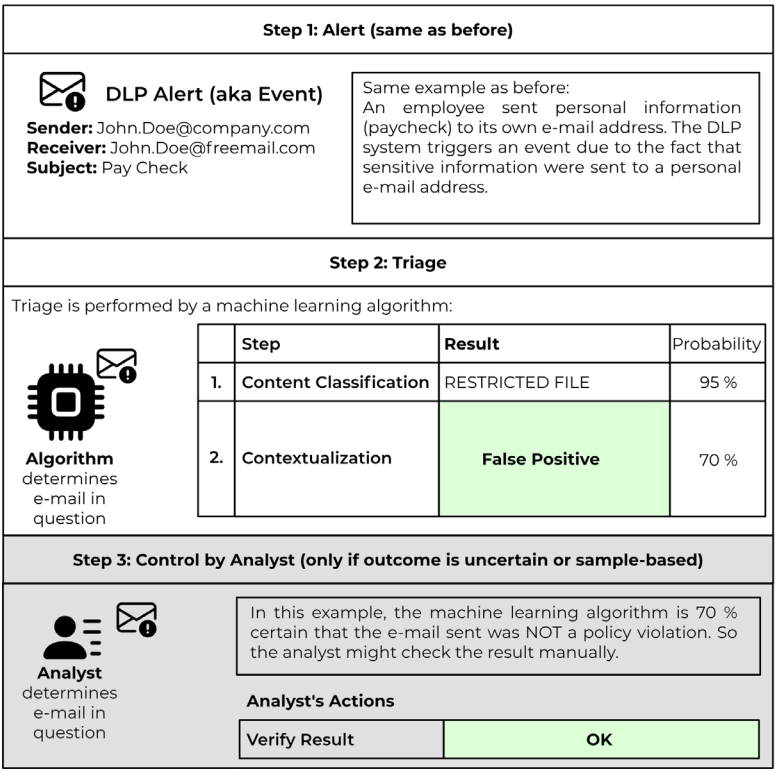

Figure 2: Research in progress - Is there an efficient way to automize the triage process like this?

6 Performance Measurement

A natural question is what would good key performance indicators (KPI's) for an automated DLP solution look like and how could they be measured against a human analyst doing the same tasks.

Regarding these questions, the general definition of the efficiency of a DLP system is determined.

Regarding the marketing slides from big vendors, the efficiency of their DLP solutions is based on a set of innovative features. However, in practice, the effectiveness of many security systems (e.g. firewalls or intrusion prevention

systems) correlates strongly with its configuration. It seems logical, that highly sophisticated and well maintained rule sets are often more effective than outdated pre-sets, provided by a vendor. Assuming that a well configured DLP environment is in place, the efficiency of the described, automated approach will be assessed by the risk rating or classification of the data that was reviewed.

In many DLP studies (e.g. (Huang et al., 2018)), algorithms are tested against test sets that are publicly available on the internet (e.g. (Lewis et al., n.d.)). This leads to indicative studies from an academic point of view. But algorithms that work on generic test sets do not necessarily work in a complex organisation, where personal - and business e-mails are often mixed up and languages vary. However, these »exceptions« are precisely where bottlenecks in modern DLP system occur, causing several false positive events as described earlier.

To give an example, newspaper articles make a brilliant test set for testing the classification skills of a machine learning algorithm. Although this of course has nothing to do with the data that will be classified by a DLP analyst who will, for example, search through suspicious web uploads with obscured, highly sensitive data.

To be able to see if an automated, contextual classification can compete with an analyst's work, it seems indispensable that a set of test data, that mimicks typical office communications, but contains clear security violations must be created, geathered and / or generated.

References

AlKilani, H., Nasereddin, M., Hadi, A., Tedmori, S., 2019. Data Exfiltration Techniques and Data Loss Prevention System, in: 2019 International Arab Conference on Information Technology (ACIT). Presented at the 2019 International Arab Conference on Information Technology (ACIT), IEEE, Al Ain, United Arab Emirates, pp. 124–127. https://doi.org/10.1109/ACIT47987.2019.8991131

Alneyadi, S., Sithirasenan, E., Muthukkumarasamy, V., 2016. A survey on data leakage prevention systems. Journal of Network and Computer Applications 62, 137–152. https://doi.org/10.1016/j.jnca.2016.01.008

Broadcom, 2022. Symantec Data Loss Prevention – Drive Total Protection of your Sensitive Data[Online]. Available: https://docs.broadcom.com/ doc/data-loss-prevention-family-en

CALO Project, 2011. Enron Email Dataset. [Online]. Available: https://www.cs.cmu.edu/~enron/

Faiz, M.F., Arshad, J., Alazab, M., Shalaginov, A., 2020. Predicting likelihood of legitimate data loss in email DLP. Future Generation Computer Systems 110, 744–757. https://doi.org/10.1016/j.future.2019.11.004

Gugelmann, D., Studerus, P., Lenders, V., Ager, B., 2015. Can Content-Based Data Loss Prevention Solutions Prevent Data Leakage in Web Traffic? IEEE Security and Privacy, vol. 13, no. 4, pp. 52–59, 2015. [Online]. Available: www.computer.org/security

Huang, X., Lu, Y., Li, D., Ma, M., 2018. A Novel Mechanism for Fast Detection of Transformed Data Leakage. IEEE Access 6, 35926–35936. https://doi.org/10.1109/ACCESS.2018.2851228

Lewis, D.D., Yang, Y., Rose, T.G., Li, F., n.d. RCV1: A New Benchmark Collection for Text Categorization Research Online]. Available: http://www.daviddlewis.com/resources/testcollections/