2009 First International Conference on Advances in Databases, Knowledge, and Data Applications

# Analyzing Malware Log Data to Support Security Information and Event Management: Some Research Results

Roland Gabriel, Tobias Hoppe,
Alexander Pastwa
Chair of Business Informatics
Ruhr-University of Bochum
Bochum, Germany
rgabriel@winf.rub.de
thoppe@winf.rub.de
apastwa@winf.rub.de

Sebastian Sowa
Institute for E-Business Security (ISEB)
Ruhr-University of Bochum
Bochum, Germany
sebastian.sowa@rub.de

***Abstract*—Enterprise information infrastructures are generally characterized by a multitude of information systems which support decision makers in fulfilling their duties. The object of information security management is the protection of these systems, whereas security information and event management (SIEM) addresses those information management tasks which focus on the short term handling of events, as well as on the long term improvement of the entire information security architectures. This is carried out based on those data which can be logged and collected within the enterprise information security infrastructure. An especially interesting type of log data is data created by anti-malware software. This paper demonstrates in the context of a project case study that data mining (DM) is a well suited approach to detect hidden patterns in malware data and thus to support SIEM.**

***Association Analysis; Cluster Analysis; Data Mining; Information Security Management; Log Data; Malware; Security Information and Event Management***

## I. Introduction

It is a well known fact in the fields of general business management research and business informatics that the efficient and effective processing of information through the use of adequate information systems constitutes an important driver for the success of an enterprise [1]. The organization's functions, as well as the internal and external processes are highly dependent on information and the information systems, which semi- or fully automatically support information processing [2].

Considering that a shortfall of essential information systems may lead to existential dangers, special attention must be paid to the availability of the information infrastructure which in this case means every device and application used for processing information. Furthermore, breaches in the confidentiality, integrity, non-repudiability in connection with information or information processing may constitute perceptible impairments or even existential crises [3]. The protection of these security objectives is therefore one of the main goals of information management (IM), that in general aims to support the enterprises' executives with the optimally designed and run information infrastructure. The tasks connected with attaining the aforementioned security objectives in this context are thereby attributed to information security management (ISM).

An integrated bundle of measures (containing organizational, technical, logical as well as physical measures) is needed for the realization of the defined security objectives [4; 5]. In this context ISM includes the steering and control of these measures, as well as their initial planning. This process must be seen as a continuous operation and implemented accordingly to guarantee a sustainable realization of the desired level of protection [6; 7]. In this regard information also plays an important role, as it forms the basis for possible modifications of the measures which further improve the desired level of protection.

As a subdivision of ISM the security information and event management (SIEM) discussed in this paper typically uses a wide range of information from various elements of the information security architecture (IS architecture). The IS architecture reflects those components of the information infrastructure which practically enforce security objectives, as well as they can be used for the administration of the relevant and underlying concepts. These elements compromise all access controls, operating system cores, firewalls and measures to guarantee safe communication [8].

As comprehensive as the amount of elements of the IS architecture is, as comprehensive is the volume of data generated from its elements either for which reason data evaluation is complex and time consuming. Therefore, a critical success factor for SIEM is the quality and not the quantity of data relevant for decisions about implementing or modifying possible measures.

978-0-7695-3550-0/09 $25.00 © 2009 IEEE
DOI 10.1109/DBKDA.2009.26

Due to the amount and the complexity of data that have to be analyzed, the question about tools, methods and models to support the analysis process arises. One approach is data mining. After dealing with the theoretical backgrounds concerning SIEM in chapter II., chapter III. focuses the research objectives of this paper. Chapter IV. deals with the proceeding of the data analysis whereas chapter V. refers to its results. Chapter VI. gives a brief conclusion and finally, chapter VII. exemplifies future prospects.

## II. Theoretical Background – SIEM

As seen in figure 1, SIEM combines security information management (SIM) and security event management (SEM). In both fields the focus lies on the collection and analysis of security relevant data. However, SEM emphasizes the aggregation of data into a manageable amount of information with the help of which security incidents can be dealt with immediately (i.e. in a timely fashion), while SIM primarily focuses on the analysis of historical data in order to improve the long term effectiveness and efficiency of information security infrastructures [9]. The amalgamation of SIM and SEM into an integrated process of planning, steering and controlling security relevant information on the basis of data collected from the IS architecture is summarized under the term SIEM [10].

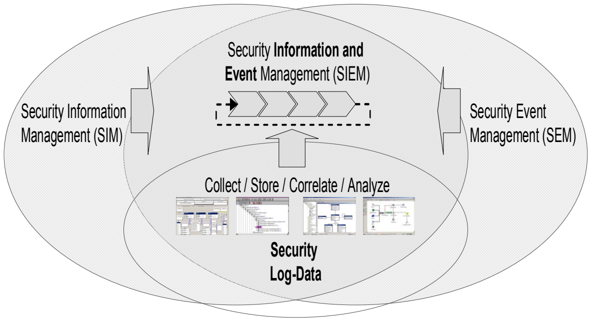
Security Information and Event Management (SIEM)
Security Information Management (SIM)
Security Event Management (SEM)
Collect / Store / Correlate / Analyze
Security Log-Data

Figure 1. Conceptual Architecture of SIEM

By identifying information and deducting knowledge from the existing volume of data SIEM strives to guarantee the protection of the information and the information system asset values of an organization. To achieve this goal it is necessary to conduct SIEM as an integrated and continuous management process. In turn, this process is dependent on information relevant to decision making which is extracted from the data pool. It is therefore crucial to establish appropriate practices and mechanisms which support the utilisation of data in the management process as effectively and efficiently as possible.

As a result of the numerous components installed in an IS architecture, the volume of protocols as well as the amount of data generated is enormous. Depending on the system and the action performed, log data may contain information about incidences or threats (due to email or internet use, for example). In addition, the data in question may be recorded because specific ports were used by gateways and firewalls [11].

## III. Research Objectives

The simple storage of security relevant data alone does not enable researchers or business analysts to draw sensible conclusions from the data in order to support SIEM. Data by itself is of little direct value since potential insights are buried within and are often very hard to uncover. The concept of data mining provides specific algorithms for data analyses like association analysis, clustering or classification [12]. The methods and algorithms which are used for these purposes originate from a variety of research fields. These fields include statistics, machine learning, pattern recognition, database research, business intelligence and data visualization.

It must be clear, that the application of data mining algorithms must be accompanied by preparatory as well as post processing steps [12]. As Fayyad et al. put it, »blind application of data mining methods can be a dangerous activity, easily leading to the discovery of meaningless and invalid patterns« [13]. In order to conduct the necessary preparatory and post processing steps as well as to analyze the data efficiently and effectively the Cross Industry Process for Data Mining (CRISP-DM) was applied [14].

Actually, every single log event is potentially interesting for investigative analysis. Since most organizational IT networks are in some way connected to the Internet and are thus subject to attacks from outside, the most popular application of data mining on log data is concerned with intrusion detection [15; 16]. In addition questions to be answered by analyzing the log data can be for example why, where, when and how long a malware incident happened and who was involved and responsible. In order to attain new and useful insights from the log data of interest, the following research objectives were identified.

### *A. Objective 1: Malware Factor Analysis*

One goal of applying data mining techniques is to identify interesting and relevant patterns in the data, which can be verbalized and quantified in form of rules. The resulting set of rules can then be further analyzed by a human expert who decides how these rules will further be used in the process of SIEM. Among the different methodologies which are used to extract rules from a given data set, we focussed on the association analysis. Association analysis aims to discover interesting relationships between the attributes of a data set [17]. Support and confidence are two measures for the interestingness of a relationship. Support indicates how frequently a rule is applicable to a given data set, while confidence determines how often items in B appear in transactions that contain A [17]. As seen in figure 2, the support of 2% means that in 2% of the whole set of hosts, Windows XP and a malware incident went along with each other. The confidence of 10% indicates that malware incidents occurred on 10% of all Windows XP hosts.

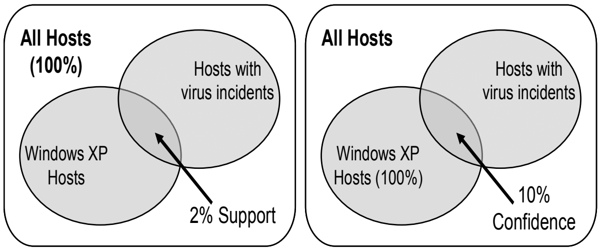

Figure 2. Support and Confidence

Mathematically, support and confidence can be represented as in the following equations, where A is the antecedent and B the consequent of the rule:

– Support $(A \rightarrow B) = P(A \cup B)$;

– Confidence $(A \rightarrow B) = P(B \mid A)$.

Since many relationships may exist between the attributes causing malware incidents the following research objective was stated: “Given malware incidents with certain attributes, find associations between those attributes and state them as rules satisfying a certain confidence and support.”

### B. *Objective 2: Malware Permanence and Propagation Analysis*

In addition, the given malware log data set can be analyzed against the background of the question of how malware spreads in the IT landscape and how long it resides in the system. Such a profile may contain data about the number of computers and users affected and the duration the malware resided in the IT infrastructure. Thus the second objective of mining the security relevant data was to enable an analysis of malware permanence and propagation.

It was found that for this objective, no usual data mining methods were applicable. Instead, a specific clustering algorithm had to be devised in order to cluster malware incident records in dependence of their similarity. Describing similarity is a main task of clustering algorithms. Similar records are put into the same cluster, whereas dissimilar records are put into different clusters. Thus, the second research objective was stated as follows: “Given a set of *n* malware incidents, group them by similarity into *k* clusters”.

## IV. Proceeding

The proceeding and the results being discussed as follows are the key findings from a cooperative project between a university and an industrial institution of leading presence. The goal was to develop a solution for a more sophisticated analysis of information security relevant data. The industrial institution uses a combination of several security systems. Their generated log data is stored in a centralized relational database. Amongst others, main sources of the log data of interest are those from anti-malware solutions.

Figure 3 illustrates the business objectives of the data mining project and the way log data contributes to them. A log event thereby is a specific, single event created by some log source and stored in the database. An example for a log event is the finding of a malware program. Log events are very numerous and hard to analyze, so they are aggregated to incidents. An incident thus covers one or more log events which belong together. One and the same malware might create multiple log events in a given timeframe (e.g. for every file it was found in), but it still could be considered the same malware on the same computer. A possible aggregation of malware events to malware incidents might be based on malware events occurred on the same computer and caused by the same user in a predefined timeframe.

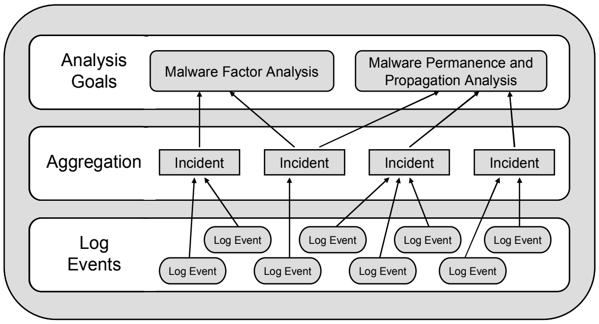

Figure 3. Aggregation of Log Events

Figure 4 gives an overview of the data made available for this case study. The data set can be separated into actual log data and context data. The actual log data is divided into three types. On the one hand logs contain log data originating from the Windows operating systems. On the other hand for Unix hosts, similar data was made available. The most interesting log data in respect to the research objectives stated above is the malware log data. The malware event records contain information about the time, location and type of malware found on a system.

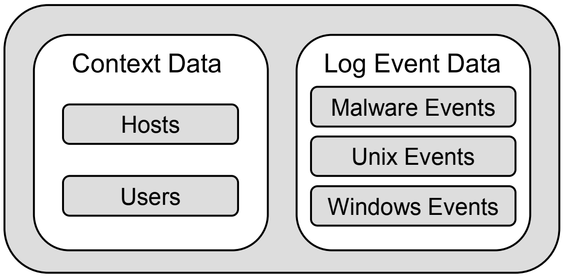

Figure 4. Overview of Input Data

The context data consists of records representing the computers (hosts) and the users of the company's IT systems. These records offer data in several dimensions such as geographic and demographic information. The user records include fields containing information like the user's age and gender as well as his or her organizational status within the company. The host records include fields containing the computer's current status and the operating system running on it as well as information about the patch status of these operating systems.

## V. Results

Since the original results of the data analysis could not be published due to confidentiality requirements, it has to be stressed that the following findings base on random data. Nevertheless the results convey an impression on the possible outcomes of such an analysis.

### A. *Findings of the Malware Factor Analysis*

Since the Apriori algorithm is applicable for analyzing small or mid-size data sets we decided to apply this algorithm to provide a solution to research objective 1 [17]. Table 1 depicts an extract of some random data on which the Apriori algorithm was applied.

TABLE I. Overview of Database Extract

| No. | User Age | User is Admin | Malware Risk |
|---|---|---|---|
| 1. | IV | true | low |
| 2. | V | false | low |
| 3. | III | false | low |
| 4. | II | true | high |
| 5. | II | false | high |

Each row represents a virus incident with 3 attributes. During the project it turned out to be useful to categorize the users' ages into different classes. The classes range from "I" for the youngest employees to "V" for the eldest employees. In order to find out which attributes are associated with high malware risk (or low malware risk, respectively) the different types of malware had to be assessed prior to the analysis. This was done by adding a new attribute to the data table for malware risks. Thus, it was possible to assign each user a "low" or "high" malware risk. Cookies, adware and joke programs were classified as low risk while malware such as viruses, trojans and key loggers was classified as high risk.

Since the case study focused the question what factors affect malware affection, only those item sets containing the risk attribute were taken into account. In order to gain significant rules, a threshold for the support and confidence had to be determined. The rules derived from the random data are listed in table 2.

TABLE II. Association Rules

| ***high* malware affection, if** | | **Support %** | **Confidence %** |
|---|---|---|---|
| 1. | user age category = IV and<br>user gender = male | 9.5 | 82.7 |
| 2. | user age category = III and<br>user gender = male and<br>user is admin = false | 5.3 | 75.6 |
| ... | ... | ... | ... |

| ***low* malware affection, if** | | **Support %** | **Confidence %** |
|---|---|---|---|
| ... | user is admin = true and<br>user gender = female | 1.5 | 50.7 |
| n. | user age category = IV and<br>user gender = female | 8.7 | 60.9 |

The application of the Apriori algorithm made it possible to separate the rule set. Rules with a confidence of less than 70% and a support of less than 5% were not taken into account. The upper part of the table displays the rules which lead to high malware affection. The lower part displays those rules with a low malware affection, respectively. The support and confidence percentages are displayed for each rule. For example the support of rule 1 allows the conclusion that in 9.5% of malware incidents the user's age category is IV, the user's gender is male and the malware affection was high. The confidence of rule 1 indicates that in 82.7% of those malware incidents where the user's age category is IV and the user's gender is male, the malware affection actually is high.

It was tempting to interpret the rules indicating low malware affection similarly. However, the analysis only included records which already represented at least one malware incident. The incidents categorized as 'low malware affection' merely occurred on hosts with not so many malware incidents. Thus, the last two rules were to be handled with care, since they merely indicated lower malware affection than e.g. rules 1 and 2, but not a complete absence of it.

### B. *Findings of the Malware Permanence and Propagation Analysis*

Data mining with the goal of describing the permanence and propagation of malware incidents throughout the hosts of the company was not performed in a straightforward fashion such as for the malware factor analysis. The efforts put into this task are described in the following.

In order to narrow the focus of the analysis, some measures for malware permanence and propagation had to be defined. The propagation of malware is best described by the number of hosts and number of users a specific malware has affected. The duration of a malware infection may serve as a measure of malware permanence. With these measures defined, the next task was finding a way to derive the measures from the data. The malware event data served as basis for this analysis, so no further data preparation was needed.

The most difficult measure to extract from the data was the duration of a malware infection. A malware infection in this context is defined as the duration in which the same malware was present on different hosts throughout the company. So, if a specific malware was identified on at least one host at the beginning of April and again in the middle of April, we are dealing with two separate infections. The malware incident data thus had to be aggregated once more to provide information about such infections. This time, the aggregation had to be performed along the date attribute of malware incidents. Malware incidents with the same malware and a similar date were to be aggregated to the same malware infection. But when is one date similar to another and when are they dissimilar? Is the first of April similar to the third of April? Is it similar to the fifth?

It seemed that the task of aggregating the malware incidents to malware infections could be solved with a clustering algorithm. The date clustering algorithm devised for the case study clusters data objects by date and malware ID. The results are a number of clusters, each containing a number of data objects with the same malware ID and a similar date.

The algorithm performs the following steps for each identified malware ID:

(1) Sort all data objects by date.

(2) Create an initial empty cluster.

(3) Go through the data objects in sorted order and compare the date of each object to the date of the previous one. If the dates are similar, put the current data object into the currently opened cluster. Otherwise, close the currently opened cluster and create a new one containing the current data object. Similarity between dates may be parameterized. In the case above, two dates were considered dissimilar if they were more than 7 days apart and otherwise were considered similar.

Finally, the attribute 'cluster' was added to each record. This attribute will have the value '0' if the record belongs to no cluster and a different number if it is part of a malware infection cluster. The result was a number of clusters, each containing a number of data objects with the same malware ID and a similar date.

The clustering algorithm was parameterized during some test runs in such a way that most clusters contain either mostly malware incidents with high malware affection or mostly those with low malware affection in order to be able to identify malware factors. The attribute distributions of the clusters indicate if the administrative privileges, the age and the gender result in uncommon malware affection. Due to confidentiality requirements the results of the cluster analysis cannot be discussed in detail.

## VI. Conclusion

While many research papers focused the analysis of log data e.g. for web marketing purposes the analysis of security-relevant log data has not been widely explored yet. As a result of the project it could be exemplified that the so called native data mining methods are applicable for the analysis of security relevant log data.

Although the results presented in this paper are based on random data, rules were identified throughout the data mining project indicating that the age of a user has impact on malware affection on the one hand and that the user's gender also influences malware occurrences on the other hand. At the same time it had to be stated that the admin status of a user does not seem to have influence on malware affection. It has to be pointed out though that these findings should not be generalized as they may relate to specific circumstances of the project conducted.

Due to the amount of data being processed throughout the project, major efforts had to be made to ensure the quality of the log data in regard to its readiness for analysis. Though not being in the focus of this paper, it has to be stated that the application of a data mining process, like CRISP-DM for instance, is a crucial success factor to achieve this goal.

## VII. Future Prospects

Naturally, the results of the malware factor analysis should provide information about which factor influenced the number of malware occurrences on the company's hosts. An easily understandable representation of such information is in the form of rules. A rule might say 'if a user has administrative privileges on a host, this host does not have an abnormally high number of malware incidents', which is completely unambiguous and easy to understand and explain. The fact that a rule is easy to understand does not mean that it must be correct, however. Such rules can be generated by an association analysis. More specifically, the Apriori algorithm was applied to analyze the malware factors.

As for research objective 1 discussed in this paper it seems sensible to create another model based upon a different technique in order to support or disprove the rules generated by Apriori. This can be achieved by training a clustering model with the k-Means algorithm. An association rule might be supported by the cluster analysis, if at least one cluster can be associated to it. A cluster representing the rule stated above might contain only those records in which the user possessed administrative privileges and the host was subject to a relatively low number of malware occurrences.

In order to serve the goals of SIEM, future research has to focus on further fields of log data analysis. For example, policy violations could also be monitored by the use of the data mining methods presented in this paper. Since companies usually have a bulk of policies to which users and hosts have to comply to, like password and access rules or the enforcement of regular updates of anti-malware and operating system software, this field of security relevant data can't be handled manually. By applying the described data mining techniques in this data context, factors for violations of policy compliance can be identified efficiently as well as countermeasures can be set up in a brief time frame in the consequence. Here, the identified policy violation issues should be categorized, rated and displayed automatically in a clearly arranged manner in order to provide the information security management executives with high-quality information.

For this purpose, using business intelligence (BI)-systems should be taken into consideration. Thereby, data mining is one possibility to identify patterns in the data which were previously hidden. Another way to perform data analyses and visualize the results can be done by looking at reports created by BI-tools which are part of a BI-system. ETL-tools which are also used in a BI-system allow the extraction, loading and transformation of security relevant data stored in relational databases into a data warehouse according to a multidimensional data model. In this context online analytical processing (OLAP) as one of the most promising technologies that are usable for BI provides flexible and fast response analyses along previously defined dimensions. Such dimensions can be the aforementioned hosts, users, different types of malware as well as security threats for instance. If also a time dimension is added to the multidimensional data model, even time series analyses are performable.

To sum up, the possibilities of BI in the context of SIEM are manifold. Thereby, data mining techniques offer the opportunity to extract new knowledge out of the seemingly unstructured set of data logged. This knowledge again enables to design new or adjust current measures resulting in an enhancement of the quality of the entire information security infrastructure of the company of interest.

## ACKNOWLEDGMENT

Sebastian Sowa thanks Roland Gabriel, Ruhr-University of Bochum, Germany, for granting a sabbatical as well as Burt Swanson, University of California, Los Angeles, USA, for the invitation to spend this sabbatical at UCLA Anderson School of Management, where he primary worked on the paper presented.

## REFERENCES

[1] K.C. Laudon, and J.P. Laudon, Management Information Systems, Managing the Digital Firm, Prentice Hall International, Upper Saddle River, 2005.

[2] J.-C. Laprie, "Dependability of Computer Systems: from Concepts to Limits", Proceedings of the 6th International Symposium on Software Reliability Engineering, 1995, pp. 2-11.

[3] S.C. Shih, and H.J. Wen, "Building E-Enterprise Security: A Business View", Information Systems Security, Vol. 12, No. 4, 2003, pp. 41-49.

[4] R. Anderson, Security Engineering, A Guide to Building Dependable Distributed Systems, Wiley & Sons, New York et al., 2008.

[5] B. Schneier, Secrets and Lies, Digital Security in a Networked World, Wiley & Sons, New York et al., 2004.

[6] International Organization for Standardization, ISO/IEC 17799:2005, Information technology – Code of practice for information security management, 2005.

[7] International Organization for Standardization, ISO/IEC 27001:2005, Information technology – Security techniques – Information security management systems – Requirements, 2005.

[8] M. Nyanchama, and P. Sop, "Enterprise Security Management: Managing Complexity", Information Systems Security, Vol. 9, No. 6, 2001, pp. 37-44.

[9] A. Williams, "Security Information and Event Management Technologies", Siliconindia, Vol. 10, No. 1, 2006, pp. 34-35.

[10] D.F. Carr, "Security Information and Event Management". Baseline, No. 47, 2005, p. 83.

[11] B. Gilmer, "Firewalls and security", Broadcast Engineering, Vol. 43, No. 8, 2001, pp. 36-37.

[12] J. Han, and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, 2006.

[13] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases", AI Magazine, 1996, pp. 37-54, URL: http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf.

[14] P. Chapman, J. Clinton, R. Kerber, T. Khazaba, T. Reinartz, C. Shearer, and R. Wirth, "CRISP-DM 1.0 Step-by-Step Data Mining Guide", URL: http://www.crisp-dm.org/CRISPWP-0800.pdf.

[15] D.G. Conorich, "Monitoring Intrusion Detection Systems: From Data to Knowledge", Information Systems Security, Vol. 13, No. 2, 2004, pp. 19-30.

[16] K. Yamanshi, J.-I. Takechu, and Y. Maruyama, "Data Mining for Security", NEC journal of advanced technology, Vol. 2, No. 1, 2004, pp. 13-18.

[17] V. Kumar, M. Steinbach, and P.-N. Tan, Introduction to Data Mining, Addison Wesley, Upper Saddle River, 2005.