Business Intelligence Based Malware Log Data Analysis as an Instrument for Security Information and Event Management

Tobias Hoppe
Chair of Business Informatics
Ruhr-University of Bochum
Bochum, Germany
thoppe@winf.rub.de

Alexander Pastwa
Steria Mummert Consulting
Dusseldorf, Germany
alexander.pastwa@steria-mummert.de

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

***Abstract*—Enterprises face various risks when trying to achieve their primary goals. In regard to the information infrastructure of an enterprise, this leads to the necessity to implement an integrated set of measures which should protect the information and information technological assets effectively and efficiently. Furthermore, tools are needed for assessing risks and the performances of measures in order to guarantee continuous effort to protect the enterprises' assets. These tools have to be able to support the handling of the vast amount of security relevant data generated within the enterprise information infrastructure and their analysis. Both tasks are typical for security information and event management. In this context, the current paper introduces an approach for malware log data analysis by using business intelligence methods. Thereby, examples are given which are derived from the results of a project being conducted with a world-wide operating enterprise.**

Business Intelligence; Data Mining; Malware; Online Analytical Processing; Security Information and Event Management

I. INTRODUCTION

In general business management research as well as in the field of business informatics, it is a well known fact that the effective as well as efficient processing of information constitutes one of the most important drivers for the success of an enterprise [DBKDA 2009, 1; 2]. For this purpose, adequate information systems are used. The organization's functions and processes highly depend on information and on those information systems, which semi- or fully automatically support information processing [3].

Considering that already a temporary unavailability of essential information systems may lead to existential dangers, special attention must be paid to measures which ensure that all devices and applications of the information infrastructure being necessary for the information processing activities are used. Furthermore, breaches in the confidentiality, integrity, and the non-repudiability in regard to information assets or information processing technologies may constitute perceptible impairments or even existential crises [4]. The protection of these security objectives therefore is one of the central goals of information management, which generally aims to support the executives with an optimally designed and run information infrastructure. Tasks and responsibilities focusing on the achievement of the aforementioned security objectives are attributed to the subdivision respectively -function of information security management.

An integrated bundle of measures (containing organizational, technical, logical as well as physical measures) is needed for the realization of the defined security objectives [5; 6]. Here, information security management includes the steering and controlling of measures as well as their initial planning. This process must be seen as a continuous operation to guarantee a sustainable realization of the desired level of protection [7; 8]. In this context, information again incorporates a very important role – it forms the basis for any possible modification of the measures aiming to hold or improve the level of protection which is defined by the executives on the basis of an analysis of threats and economic impacts.

As subdivision or sub-function of the information security management of an enterprise, 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. The information security architecture is defined as the part of the information infrastructure which contains all components to enforce the defined information security objectives. Further more, these components can be used for the management and re-engineering of the relevant security concepts. From this background, the architectural elements compromise all access controls, operating system cores, firewalls and further measures to guarantee safe communication, for instance [9].

As comprehensive as the amount of elements of the information security architecture is, as comprehensive is the amount of data generated from its elements. As consequence, the task of data evaluation is complex and time consuming. Therefore, a critical success factor for executives of SIEM has to be seen in the quality and not

the quantity of data relevant for the decisions about the conceivable modifications of security measures.

Due to the amount and complexity of data that have to be analyzed, questions about adequate tools, methods and models to support the analysis process arise. Here, one of the most successful applied approaches in the business management context is business intelligence (BI). This paper shows how BI can be used to answer two questions which are relevant for SIEM: 1. How do malware causing attributes relate to each other? 2. How does malware spread in the IT landscape and how long does it reside in the system? For these purposes, known malware which occurred within a certain timeframe will be analyzed.

After dealing with the theoretical backgrounds concerning SIEM in Chapter II, Chapter III introduces the concept of business intelligence. Chapter IV shows how Online Analytical Processing (OLAP) can be applied for SIEM. Chapter V then focuses the research objectives of this paper from the perspective of data mining whereas Chapter VI refers to its results. Chapter VII gives a brief conclusion and finally, Chapter VIII exemplifies future work.

II. THEORETICAL BACKGROUND – SIEM

Before presenting how BI, in particular OLAP and data mining, may support the goals of SIEM, the following paragraphs characterize specific problems of data analysis as well as the requirements for designing a BI system. In the first step, terms and definitions which are relevant for the overall conceptual coherences are introduced.

A. Relevant Terms and Definitions

Information as the first relevant term used in the discussion of information security management topics can linguistically be derived from the Latin informatio. In this turn, informatio stands for the explanation or interpretation of ideas as well as it can be used in the meaning of education, training or instruction. This gives a first consideration about an accurate and precise definition: Information in this paper is defined as an explanatory, significant assertion that is part of the overall knowledge as well as it is seen as specific, from human beings interpreted technical or non-technical processed data [10; 11].

The just given definition of information is precisely in line with the ISO/IEC standards which explain that information "can exist in many forms. It can be printed or written on paper, stored electronically, transmitted by post or by using electronic means, shown on films, or spoken in conversation" [7; 8]. This – mostly trivial – way to use the term information unfortunately does not reflect the common sense in the information security community. There, it is quite often assumed to only affect electronic data, and thereby information security management has mostly to deal with IT. Although this paper focuses on data gathered from technological elements, it is stressed that this only covers one aspect of the entire tasks of information security management executives.

As consequence of the appreciation of information, also information security has to cover technical as well as non-technical challenges. In this context, the ISO explains that whatever "form the information takes, or means by which it is shared or stored, it should always be appropriately protected. Information security is the protection of information from a wide range of threats in order to ensure business continuity, minimize business risk, and maximize return on investments and business opportunities" [7; 8].

The term SIEM combines security information management and security event management. In both areas, the focus lies on the collection and analysis of security relevant data in information infrastructures respectively the security infrastructures. Thereby, the security event management emphasizes the aggregation of data into a manageable amount of information in order to deal with events and incidents immediately (for example, in a timely fashion).

In contrast to security event management, security information management primarily focuses on the analysis of historical data aiming to improve the long term effectiveness/efficiency of the information security infrastructure [12].

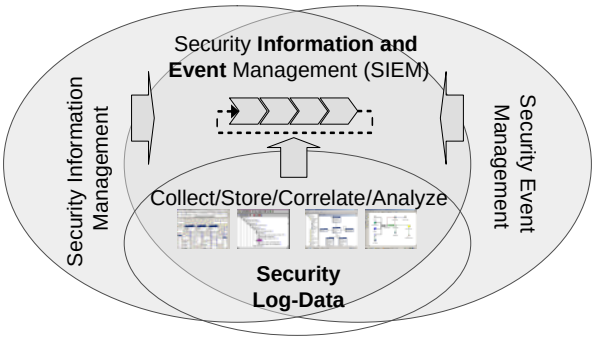

Figure 1. Conceptual Architecture of SIEM

As shown in Figure 1, SIEM then stands for the amalgamation of security information management and security event management into an integrated process of planning, steering, and controlling security relevant information on the basis of the data collected from the information security architecture. Carr states: "Security information and event management (SIEM) systems help to gather, store, correlate and analyze security log data from many different information systems" [13].

B. Selected Challenges SIEM is facing

The analysis of security relevant data collected from the information security architecture is a challenging task because of the following reasons:

- Amount of data
- Heterogeneity of data formats
- Heterogeneity of the data contents
- Limited personnel and budget

As consequence of the various information security architecture elements and the number of protocols, the amount of data gathered is massive. Thus, considerable manual effort is needed to gather relevant information about security threats. Furthermore, the data collected exist in various formats, making evaluation difficult and time-consuming. The heterogeneity of the data contents also impedes a simple and flexible analysis. Depending on the system and the action performed, the data may contain information about incidents or threats due to email or internet use, for example. In addition, data may be recorded, since specific ports are used by gateways and firewalls, for instance. Therefore, the possibility of manually analyzing data which are derived from the information security architecture elements is severely limited due to the sheer volume of data as well as the heterogeneity of data formats and contents.

Two further aspects must be considered. Typically, information security management divisions have only a small fraction of personnel, and the budget is also limited. As well as in other entities of an enterprise, the resources also spent for SIEM have to be managed economically. Thus, SIEM faces the same requirements as the other organizational units of the entire enterprise. The executives have to allocate resources in such a way that the specific entity contributes to the enterprise's goals as much as possible [14]. To sum up, the following aspects are identified as the primary requirements for SIEM:

- Extraction of information and knowledge
- Establishment of an integrated and continuous management process
- Effective and efficient data evaluation
- Support for network management
- Support for compliance management

By identifying relevant information and deducting knowledge from the existing volume of data, SIEM strives to guarantee the protection of information and information system values. To achieve this goal, it is necessary to conduct SIEM as an integrated, continuous management process. In turn, this process is dependent on the information relevant to the decision makers. This information again is extracted from the data pool. From the background of the limitations of data evaluation as described above, it is crucial to establish appropriate (what means highly effective and efficient) practices and mechanisms to support the data processing for the needs of the SIEM executives.

As consequence of the numerous elements installed in the enterprise information security architectures, the number 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 relevant for security information and event management (SIEM) may be recorded because specific ports were used by gateways and firewalls, for instance [15].

III. The Concept of Business Intelligence for Supporting SIEM

After describing the challenges of SIEM, the current chapter focuses on the introduction of the concept of business intelligence (BI).

Business intelligence stands for a conceptual framework which bundles numerous approaches, tools and applications used for the analysis of business relevant data [16]. The general aim of BI is to support effective and efficient business decision making for what purpose a data warehouse is built up. Usually a data warehouse serves as the central storage system of a BI system. For implementing a BI application serving the goals of SIEM, a reference architecture has to be defined initially. Here, Figure 2 shows the layers and elements of an architecture that serves as a basic guiding topology in this context.

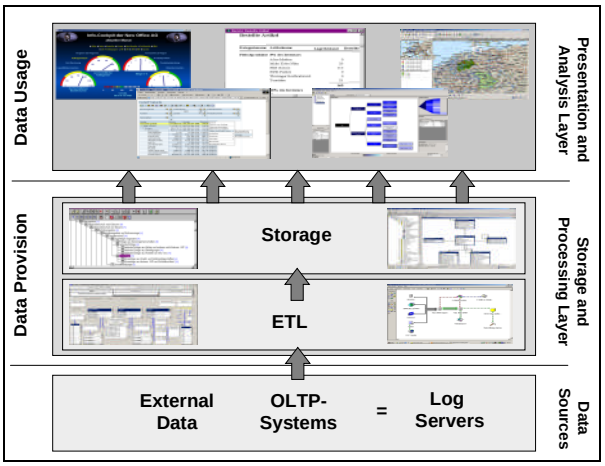

Figure 2. BI Reference Architecture [17]

A. Data Sources

At the lowest level of the BI reference architecture, various enterprises' operational systems as well as useful external data sources are located. They serve as data suppliers for the data warehouse as the integral part of the middle layer. The data primarily relevant for SIEM is gathered from the information security architecture elements which log information security relevant processes and incidents. This includes data about installed operating systems, versions of patches, installed anti-malware programs or information about the frequency of user password changes, for instance.

Potential threats can be identified by logging policy violations, malware reports, login-/logout-events and account-lockouts of users. This data is transported to log servers providing the input data for the data warehouse.

B. Storage and Processing Layer

One goal of a BI application is the consolidation of different data contents and formats towards a uniform perspective. For this task, an ETL (extraction, transformation, and load) component is combined with solutions for storing and preparing the data for later presentation / analysis [17]. This component constitutes a further module of the BI architecture and serves as the interface between the operational systems and the data warehouse [18]. It transfers the heterogeneous data into a consistent and multidimensional data perspective and loads the data into the data warehouse; in detail:

Extraction: Extraction deals with the selection and deployment of source data. Since relevant data typically exist in a very heterogeneous form, the ETL tool needs to access all data from the operational systems containing the security relevant log data.

Transformation: Transforming the source data into the target formats of the data warehouse is the central task of the ETL process. It can be further divided into the steps of filtering, harmonization, aggregation and enrichment. Filtering ensures that only the data necessary for the multidimensional analysis is loaded into the data warehouse. Log files usually contain lots of information not needed for analysis. For example, Windows event logs record a multitude of application and system information. But for the purposes of SIEM, only information security events are needed. Following, harmonization corrects the data of syntactical and semantic defects. Also an adjustment of codes, synonyms and homonyms as well as the unification of different definitions of terms will be conducted. For example, for the same person, a different user name could have been assigned in a Windows environment and in a UNIX or a Linux environment. In the multidimensional database, this user must be clearly identifiable, however. In a further step of transformation, the consistent, but in the lowest level of granularity existing data will be aggregated to improve analysis performance. Here, the aggregation of hosts to organizational units or geographical locations is a possibility. Enhancing the data by adding contextual information represents the last and very important step of the transformation process because the knowledge generated in the consequence enables to systematically substantiate decision making processes on a broader base.

Load: Finally, the extracted and transformed data is loaded into the data warehouse where it is permanently stored. For this purpose, batches are used. In order to ensure the adequate supply of information in regard to timeliness and quality, the question has to be answered how long the interval between the single batches should be. Thus, depending on the amount of data as well as on the information and communication technologies in use and the information needed by the decision makers, the data is transferred flexibly from the source systems into the data warehouse.

C. Presentation and Analysis Layer

The top layer of the BI reference architecture comprises all methods and tools which are capable to analyze the multidimensional data as well as to present analytical reports. Among the different possibilities in this context, OLAP and data mining methods play an especially prominent role:

Online Analytical Processing: OLAP is a software technology. It allows decision makers to accomplish fast, interactive and flexible requests to the relevant data stored in a multidimensional structure [19].

Data Mining: While OLAP focuses mainly on historical analysis, data mining is concerned with a prospective analysis. By applying various statistical and mathematical methods, data miners aim to identify so far unknown data patterns [20].

OLAP and data mining increase the prospect of analyzing security relevant data efficiently for the short term treatment (e.g., of malware threats) as well as for the long term improvement of the overall information security architecture. Especially in regard to the SIEM challenges, BI offers the chance to handle the accrued amount of data and to transfer the heterogeneous data into a consistent format that can be used for analyses and reports of SIEM relevant topics.

IV. Applying Online Analytical Processing for SIEM

Up to now, challenges of SIEM and characteristics of BI have been described. The following chapters focus on the combination of these fields, presenting an OLAP application for SIEM.

A. Multidimensional Data Model

Modeling an adequate multidimensional data structure is one of the crucial factors of success when designing a BI application. It forms the basis for the execution of the ETL process with which relevant data is loaded from the operational systems into the data warehouse. The resulting data construct can then be analyzed by typical OLAP operations: Slice, dice, drill down, and roll up. By using these operations, diverse occurrences of different perspectives can be determined and evaluated, like the frequency of malware infections within a certain period on a certain operating system, for instance. Figure 3 visualizes the arrangement of the dimensions mentioned above in the structure of a so called data cube [21].

Multidimensional data models consist of a fact table and further tables which serve to depict the so called dimensions. Dimensions stand for the relevant entities

with which the metrics of the fact table can be analyzed [22]. Hence, dimensions are used to provide additional perspectives to a given fact [23]. In order to ensure the data quality, it is of vital importance to follow a systematic and holistic approach when defining the dimensions and selecting the facts.

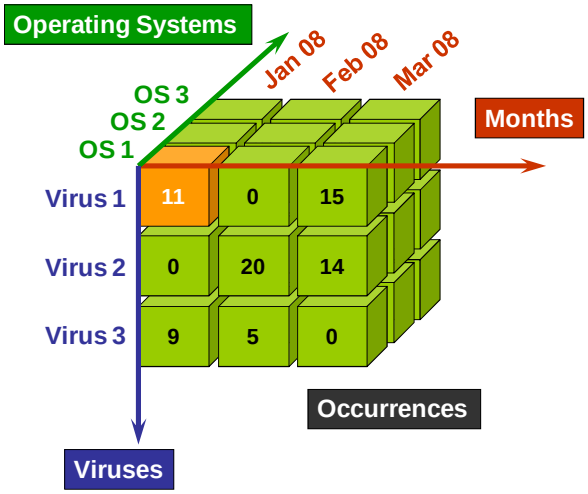

Figure 3. Example of a data cube for OLAP analyses

The content proposed in this paper refers to the key findings resulting 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. The 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 4 illustrates the business objectives of the business intelligence project and the way log data contributes to them.

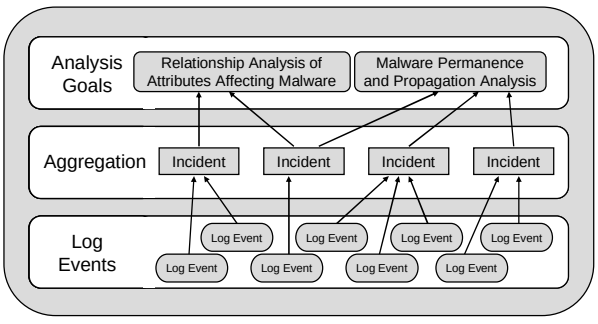

Figure 4. Aggregation of Log Events

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.

In order to aggregate malware logs, two cases must be considered:

(1) A malware which is detected at t_1 reappears on the same computer at t_2 and thus generates a new log file. For this case, the reappearance of the malware at t_2 is treated then as a new incident, if the malware has been deleted successfully in t_1 and the subsequent scan has not revealed a persistence of the malware. In addition, the malware events must have occurred on the same computer and must be caused by the same user.

(2) Further on, each log event indicating that a new malware has been detected on a computer becomes part of a new malware incident.

Figure 5 gives an overview of the input log data made available for the case study. Only known malware was in the focus of the upcoming analysis.

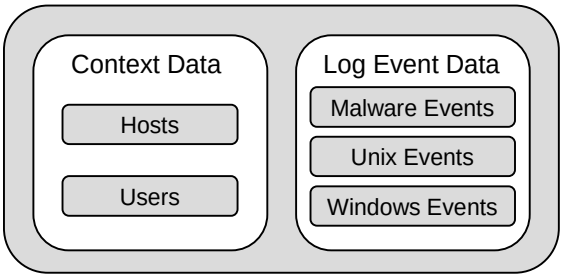

Figure 5. Overview of Input Data

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 paper is the malware log data. The malware event records contain information about the time, location, and type of malware found on a system.

The context data consists of records representing the computers (hosts) and the users of the enterprise'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 the operating system.

The resulting multidimensional data model, presented in Figure 6, illustrates the relations between the relevant dimensions containing different levels of hierarchy and the measures (facts).

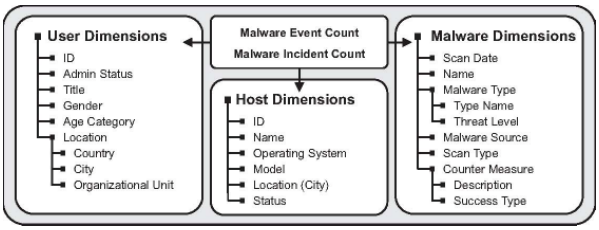

Figure 6. Multidimensional Data Model for Malware Analysis

The metrics Malware Event Count and Malware Incident Count can be analyzed according to these dimensions in any combination.

The User Dimensions include demographic information about users (e.g., gender, age category) who caused malware events as well as their admin status on the host where the malware was found. Additionally, their geographic location is tracked by the Location dimension, which consists of the hierarchy levels Country, City and Organizational Unit.

Information about the host computers on which malware events were found is provided by the Host Dimensions. The host Model is a description of the hardware. The host Status provides information about the host's current status in regard to anti-malware logging as well as information about the patch status of the operating systems.

The Malware Dimensions provide information about the Type of malware found and its Threat Level, which is either "high" or "low" by previous definition. E.g., cookies, adware, and joke programs are classified as low risks while malware such as viruses, trojans, and key loggers represent high risks. The Malware Source indicates the location of the malware; e.g., "local hard drive" or "internet browser files". Since anti-malware programs scan on a regular basis as well as on file access, the corresponding scan types are the elements of the dimension Scan Type. The countermeasures which are taken by the anti-malware software constitute the definition of another dimension (Counter Measure).

This multidimensional processed data also serve as data basis for the upcoming data mining process.

B. Prototyping an OLAP System for SIEM

Dashboards are usually used to visualize different, distributed information in a concentrated and integrated form. Relevant information is qualified in order to represent large quantities of information to the decision makers more clearly. Dashboards enable organizations to measure, monitor, and manage business objectives more effectively in the consequence [24]. In the case of SIEM, security dashboards are deployed in order to visualize security relevant data.

The dashboard illustrated in Figure 7 is currently set up to enable analysis of malware permanence and propagation. Here, the four reports merely provide descriptions of the data, indicating irregularities. Thus, they provide the starting point for a more accurate analysis, which is only possible within the individual organizational context. Since the original results of the data analysis are not allowed to be published due to confidentiality requirements, it has to be stressed that the following findings base on generated random data. Nevertheless, the results convey an impression about the possible outcomes of such an analysis.

Report no. 1 depicts the top five malware programs measured by the number of affected hosts, the number of affected users, and the duration of the malware in the institutions' IT systems. The malware "JS/Downloader-AUD" stands out, infecting 664 hosts and 431 users. It was present on at least one host on 322 days which is virtually every day in the given time frame of one year. This result implies that this particular malware either remains on the system or returns frequently.

A specific top five list of malware affections is helpful to identify particular pertinent malware and thus is a valuable tool for risk management. The types of malware visualized in the diagram can be filtered while the time period of the collected data can be adapted to one's need. The variability of such dimensions is a main feature of multidimensional OLAP analysis.

Report no. 2 illustrates the long-term development of the number of hosts and users infected with malware. Once countermeasures have been applied, this diagram can be used to control the measure effects. Scaling from quarters to months or even days, the diagram can also serve for medium to short-term controlling tasks and is thus another useful tool for risk management.

The reports no. 3 and 4 give details about the most frequent malware, in this case of the "JS/Downloader-AUD". The left diagram represents the success of malware elimination over time, the right one shows the presence of the malware in the IT systems over time. In this chart, strong excursions are to be recognized. Even after deleting the malware successfully, it seems that the malware re-emerges quickly. Further investigations concerning this malware should be accomplished.

During the project, several more dashboards were developed to enable users to analyze malware findings in regard to geographical aspects, for instance. The associated reports are represented as color coded maps in which significant occurrences of malware affection can be recognized rapidly. Further more, occurrences can be examined in detail by drilling down. With this opportunity, enterprises are able to identify locations which particularly cause the malware spreading. Thus, it can be derived in which organizational units security measures have to be improved immediately. Another dashboard visualizes user groups which cause various malware, by demographic characteristics. In this way, various age groups and/or gender-specific classes can be identified that correlate with increased malware affection. This information could be utilized to design specifically targeted awareness measures aiming to significantly reduce malware infections amongst the users and for other purposes.

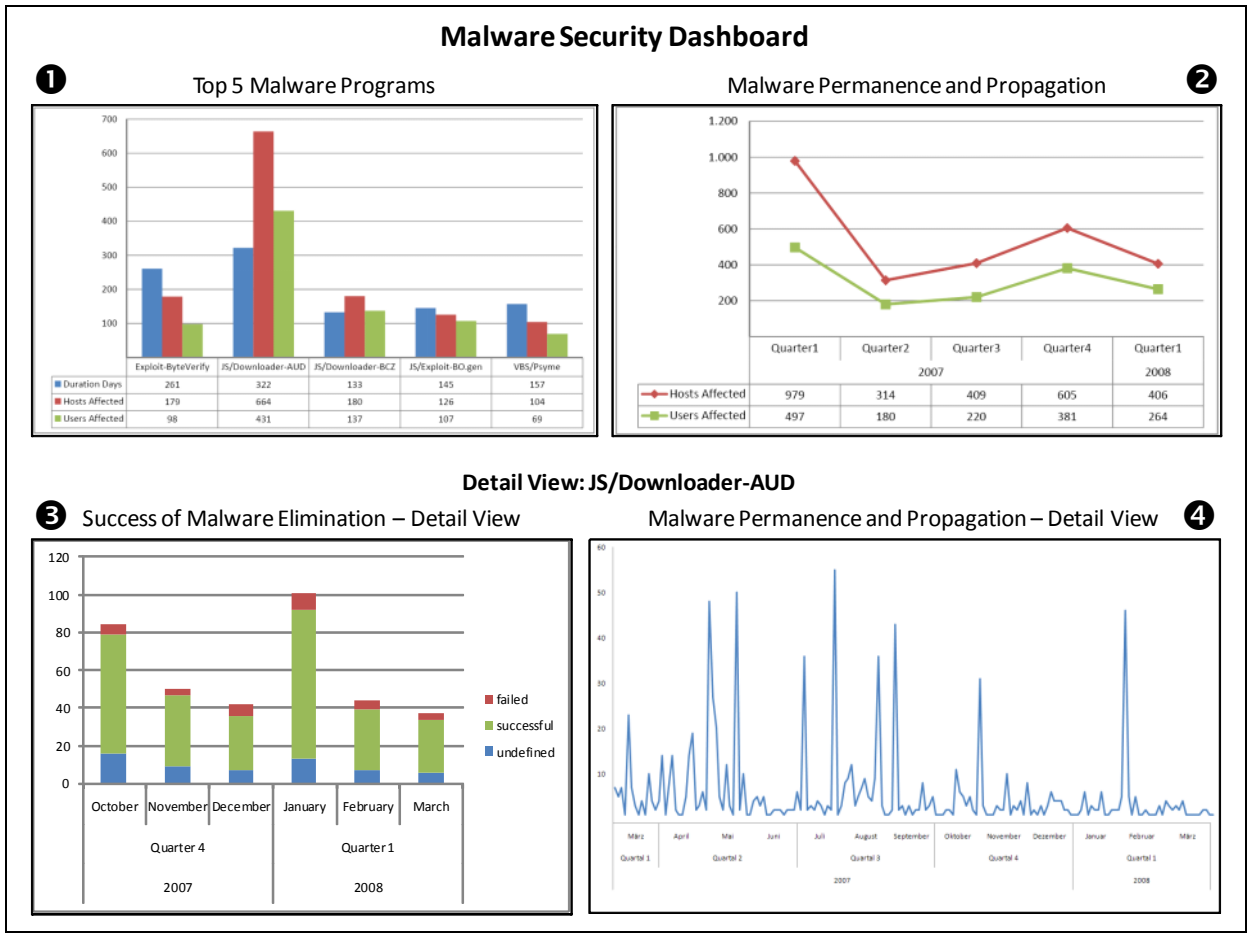

Figure 7. Excample of a Malware Security Dashboard

To sum up, all OLAP functions specified above can be used for detailed analysis in dashboards. First of all, dashboards give a general overview of the relevant measures, but also can be designed for presenting important details. Additionally, using a reporting tool, many other OLAP reports can easily be generated by accessing the data warehouse. Here, dimensions can be combined flexibly in order to analyze measures in regard to the perspectives of individual interest. In the consequence, OLAP enables a powerful descriptive analysis and effectively supports SIEM.

V. Data Mining Research Objectives

Undoubtedly, the simple storage of security relevant data alone does not enable 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. As described above, OLAP and dashboards are one way to analyze and visualize data which is modeled multidimensionally and stored in a data warehouse. Data mining is another option. The concept of data mining provides specific algorithms for data analysis, like association analysis, clustering, or classification [25]. These algorithms originate from diverse research fields, like statistics, pattern recognition, database engineering, and data visualization, for instance.

It has to be stressed, that the application of data mining algorithms must be accompanied by preparatory as well as post processing steps [25]. As Fayyad et al. point out, “blind application of data mining methods can be a dangerous activity, easily leading to the discovery of meaningless and invalid patterns” [20]. In order to conduct the necessary steps, and to analyze the data efficiently / effectively, the Cross Industry Process for Data Mining (CRISP-DM) was used [26]. CRISP-DM is an industry- and tool-neutral process model for data mining analysis which was and still is applied in several industry sectors successfully.

Actually, every single log event is potentially interesting for further 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 [27; 28]. In addition, questions to be answered by analyzing the

log data could be 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: Relationship Analysis of Attributes Affecting Malware Infection

One goal of applying data mining techniques is to identify interesting, unknown and relevant patterns in the data. Rules help to verbalize and quantify the patterns. 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, the authors of this paper focused on the association analysis. This method aims to discover interesting relationships between the attributes of a data set [29]. For this purpose, the two measures support and confidence are used. They indicate the interestingness of a relationship. Support quantifies how frequently a rule is applicable to a given data set, while confidence indicates how often items in B appear in transactions that contain A [29]. As depicted in Figure 8, 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% conveys that malware incidents occurred on 10% of all Windows XP hosts.

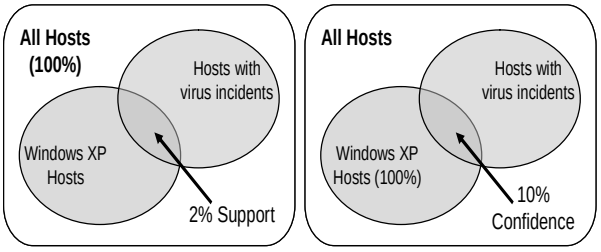

Figure 8. Support and Confidence

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 has been stated: "Given malware incidents with certain attributes, find associations between those attributes, and state them as rules satisfying a minimum confidence and support."

B. Objective 2: Malware Permanence and Propagation Analysis

Another interesting question is 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 by malware incidents and the duration the malware resides within the IT infrastructure. Thus, the second objective of mining the security relevant data aims to analyze malware permanence and propagation.

Here, the k-means algorithm was applicable in order to cluster malware incident records in dependence of their similarity. Describing similarity is the main task of clustering algorithms. Similar records are put into the same cluster, whereas dissimilar records are allocated to different clusters. Thus, the second research objective was stated as "given a set of *n* malware incidents, group them by similarity into *k* clusters".

VI. Data Mining Results

Since the results of the data mining analysis are not allowed to be published due to confidentiality reasons, the following findings also base on randomized data. Nevertheless, the results convey an impression on the possible outcomes of using data mining techniques for supporting SIEM.

A. Findings of the Relationship Analysis

Since the Apriori algorithm is appropriate for analyzing small or mid-size data sets, the authors have decided to apply this algorithm to provide an answer for research objective 1 [29]. Table I depicts an extract of random data which served as input in this context.

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
n.	...	...	...

Each row represents a virus incident with three attributes. Thereby, the user ages are grouped into one of five classes with "I" for the youngest employees to "V" for the eldest ones. In order to find out which attributes are associated with high malware risks (or low malware risks, 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. Like done for OLAP, 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 data mining analysis focused malware affecting indicators, only those item sets were regarded which contain the risk attribute. In order to gain significant rules, support and confidence factors, as shown in Table II, were calculated.

TABLE II. ASSOCIATION RULES

***hoher* Malware-Befall, wenn**		**Support %**	**Confidence %**
1.	user age category = E and user gender = male	9.5	82.7
2.	user age category = D and user gender = male and user is admin = false	5.3	75.6
...	...	...	...

***niedriger* Malware-Befall, wenn**		**Support %**	**Confidence %**
...	user is admin = true and user gender = female	1.5	50.7
n.	user age category = E and user gender = female	8.7	60.9

The Apriori algorithm made it possible to separate the rule set. Rules with a confidence of less than 70% and a support of below 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 low malware affection, respectively. The support of rule 1, as shown in the table, allows to conclude 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 age category is IV and the user's gender is male, the malware affection 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 incident. The "low malware affection" incidents merely occurred on hosts with less malware incidents. Thus, the last two rules have to be interpreted with specific attention, since they merely indicated lower affections than rules 1 and 2, for instance, but not a complete absence of it.

B. Findings of the Malware Permanence and Propagation Analysis

Data mining aiming to describe the permanence and propagation of malware incidents throughout the hosts of the enterprise was not performed in a straightforward fashion such as for the association analysis. The efforts put into this task are described now.

In order to narrow the analysis focus, measures for malware permanence and propagation were defined. The propagation of malware is described by the number of hosts and number of users a specific malware has affected. The duration of a malware infection can serve as measure for malware permanence. With background of these measures, concrete data sources were defined. Here, the malware event data served as basis for what reason no further data preparation was necessary.

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 within the entire enterprise. So, if a specific malware was identified on at least one host at the beginning of April and again in the middle of April, one is dealing with two separate infections. The malware incident data thus was aggregated once more to provide information about such infections. This time, the aggregation had to be performed along the date attribute of the malware incidents. Incidents with the same malware and similar dates were aggregated to the same malware infection group.

In order to identify similar dates, a grouping algorithm was applied. The algorithm devised for the present use case groups data objects by date and malware ID. The results were a number of classes, 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 group.
(3) Go through the data objects systematically and compare each date to the date of the previous one. If dates are similar, put the current data object into the just opened group. Otherwise, close the open group and create a new one containing the current object. Similarities between dates may be parameterized. In the case above, dates were considered dissimilar if they were more than 7 days apart.

Finally, the attribute "group" was added to each record. This attribute will have the value "0" if the record belongs to no group and a different number if it is part of a malware infection group. The result was a number of groups, each containing data objects with the same malware ID and a similar date. The grouping algorithm was parameterized during test runs in such a way that most groups contain either mostly malware incidents with high malware affection or mostly those with low malware affection.

After pre-processing the data, a cluster analysis was performed. Some findings are depicted in Figure 9. Due to the already mentioned confidentiality reason, real values must not be shown; hence, the results of the analysis cannot be discussed in detail. Since the k-means algorithm has been proven to be effective in producing good clustering results for many practical applications, this method was applied for clustering the malware incidents [30]. The attribute distributions indicate if the administrative privileges, the age, and the gender result in uncommon malware affection.

In total, eight clusters were identified. Figure 9 shows cluster 1 and 3 which were the most extensive ones. Cluster 1 includes 22%, whereas cluster 3 contains 19% of all malware incidents. Cluster 1 reveals that male users (cell 2) are likely to be affected by low-risk malware (cell 1) while in cluster 3 female users (cell 6) are in danger of being affected by high-risk malware (cell 5). Further on, cluster 1 indicates that middle-aged employees tend to be infected by malware

(cell 3). The admin status does not seem to have influence on malware infection in this cluster (cell 4) what is surprising. This is also the case in cluster 3 (cell 8). In contrast to cluster 1, cluster 3 reveals that younger and elder employees tend to have malware on their computers (cell 7).

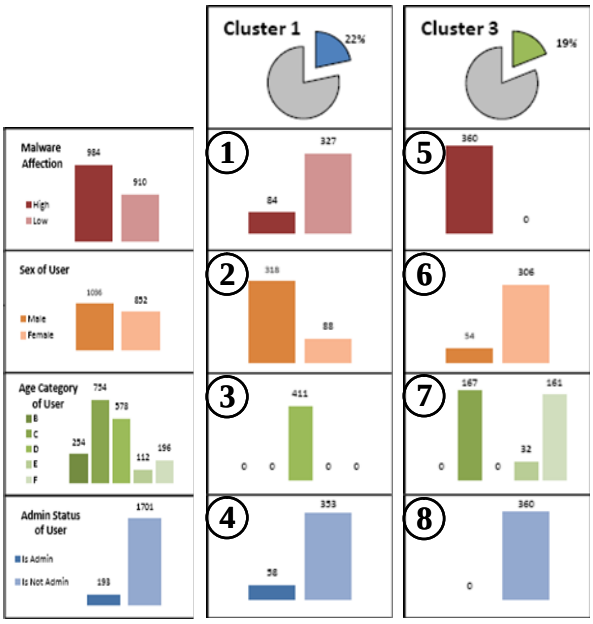

Figure 9. Results of the Cluster Analysis

VII. CONCLUSION

While BI systems are used in many enterprises to support classical business entities like the controlling or production one, they usually have little to no experience with BI systems in the context of SIEM. Taking the benefits of a classic BI system into account, this paper focused on the option of using OLAP and data mining techniques for the purposes of SIEM. Based on results of a project with an international enterprise, it can be derived that OLAP and data mining strongly support information security management teams. The gathered data can be analyzed more efficiently and patterns can be found which were previously hidden. Although the methods do not increase the detection ratio of malware directly, they support in finding internal (and external) factors which influence malware infestation. As result, measures (like awareness campaigns) could be set up to increase performances of running traditional measures like anti-virus and intrusion detection systems.

It has to be stressed that the quality of the data is crucial for success and that interpretation questions in regard to false positives and false negatives were not in the focus of this paper. Thus, the implementation of an adequate ETL process to transfer data from the source into the data warehouse correctly and consistently is as important as the validation of the accuracy of the data. In order to narrow the entire set of data to a manageable subset and to ensure that this subset matches the needs of the decision makers, the data relevance must be judged. In addition, an appropriate multidimensional data model which serves as the basis for flexible data analyses has to be designed.

While many research papers focused the analysis of log data e.g., for web marketing purposes, the analysis of security relevant log data has barely been explored. As result of the named project, it was 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 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. However, the findings should not be generalized as they may relate to specific circumstances of the project conducted.

Due to the amount of data processed during the timeframe of 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 in this context.

VIII. FUTURE WORK

Naturally, the results of the association analysis should provide information about relationships between the different attributes which influence the number of malware occurrences on the enterprise's hosts. Easily understandable representations of such information are rules. A rule might say that "if a user has administrative privileges on a host, this host does not have an abnormal high number of malware incidents".

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 be monitored by the use of data mining methods. Since enterprises usually have a bulk of policies (like password and access rules or the enforcement of regular updates of anti-malware and operating system software) to which the users and

hosts have to comply to, the corresponding security data cannot be handled manually. By applying the described data mining techniques here, factors for violations of policy compliance could be identified efficiently as well as countermeasures could be set up in a timely fashion in the consequence. Thereby, identified policy violation issues should be categorized, rated, and visualized automatically in a clearly arranged manner. Thus, the information security management executives can be provided with high-quality information. Thereby, data mining is a promising option to identify patterns inside the data sets which were previously hidden.

Another way to perform data analyses and visualize the results is OLAP. This technology leads to efficient identifications of policy compliance violations for which corresponding countermeasures could be set up rapidly. The presented OLAP approach should not only be limited to the own enterprise. Also, the standard reporting modules of anti-malware software can be substantially improved by integrating a function which enables to use dashboards as presented in the paper.

To sum up, the possibilities of BI in the context of SIEM are manifold. Thereby, data mining techniques offer the promising chance to extract new knowledge out of the seemingly unstructured set of continuously logged data on the one hand. On the other hand, OLAP enables various powerful descriptive analyses of measures according to different perspectives of interest. 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 enterprise using BI for SIEM.

REFERENCES

[1] R. Gabriel, T. Hoppe, A. Pastwa, and S. Sowa, “Analyzing Malware Log Data to Support Security Information and Event Management: Some Research Results”, Proc. First International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA 2009), IEEE Press, Mar. 2009, pp. 108-113, doi: 10.1109/DBKDA.2009.26.

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

[3] 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.

[4] 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.

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

[6] B. Schneier, Secrets and Lies, Wiley & Sons, New York et al., 2004.

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

[8] ISO/IEC 27001:2005, Information technology – Security techniques – Information security management systems – Requirements, 2005.

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

[10] J. Biethahn, H. Mucksch, and W. Ruf, Ganzheitliches Informationsmanagement, Band I, 5th Edition, Oldenbourg, München et al., 2000.

[11] R. Gabriel and D. Beier, Informationsmanagment in Organisationen, Kohlhammer, Stuttgart, 2003.

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

[13] D.F. Carr, “Security Information and Event Management”. Baseline, No. 47, 2005, p. 83.

[14] D. Hellriegel, S.E. Jackson, and J.W. Slocum, Management, South-Western College Publishing, Ohio, 1999.

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

[16] M. Anandrarajan, A. Anandrarajan, and C.R. Srinivasan, Business Intelligence Techniques, Springer, Berlin et al., 2004.

[17] P. Gluchowski and H.G. Kemper, “Quo Vadis Business Intelligence? Aktuelle Konzepte und Entwicklungstrends”, BI Spektrum, Vol. 1, No. 1, 2006, pp. 12-19.

[18] W.H. Inmon, Building the Data Warehouse, Wiley, New York et al., 1996.

[19] E.F. Codd, S.B. Codd, and C.T. Salley, Providing OLAP to User Analysts, An IT Mandate, White Paper, s.l., 1993.

[20] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From Data Mining to Knowledge Discovery in Databases”, AI Magazine, Vol. 17, No. 3, 1996, pp. 37-54.

[21] M. Jarke, M. Lenzerini, Y. Vassiliou, and P. Vassiliadis, Fundamentals of Data Warehouses, Springer, Berlin et al., 2000.

[22] W.H. Inmon, J.A. Zachman, and J.G. Geiger, Data Stores, Data Warehousing and the Zachman Framework, McGraw-Hill, New York, 1997.

[23] P. Rob and C. Coronel, Database Systems: Design, Implementation, and Management, Boston, 2007.

[24] W.W. Eckerson, Performance Daschboards: Measuring, Monitoring, and Managing Your Business, Wiley & Sons, New York et al., 2006.

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

[26] 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”, 2000, URL: http://www.crisp-dm.org/CRISPWP-0800.pdf, 22.09.2009.

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

[28] 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.

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

[30] K. Alsabti, S. Ranka, and V. Singh, “An Efficient K-Means Clustering Algorithm”, 1998, URL: http://www.cs.utexas.edu/~kuipers/readings/Alsabti-hpdm-98.pdf, 22.09.2009.