VKE: a Visual Analytics Tool for CyberSecurity Data

Antonio Gonzalez-Torres*  
School of Computing,  
ULACIT  
San José, Costa Rica  
agonzalez@ulacit.ac.cr  
http://www.ulacit.ac.cr  
and  
Department of Computer Engineering,  
Costa Rica Institute of Technology,  
Cartago, Costa Rica  
antonio.gonzalez@tec.ac.cr  
http://www.tec.ac.cr

Vetria L. Byrd and Paul Parsons  
Department of Computer Graphics Technology,  
Purdue University,  
West Lafayette, Indiana  
[vlbyrd, parsonsp]@purdue.edu  
http://polytechnic.purdue.edu

RESUMEN

Several changes have occurred throughout the past few decades regarding how organizations and individuals generate, process, and share information. The emergence of new devices has enhanced the abilities of individuals and organizations to create unique and different types of data, which are shared in real time at high speeds. Thus, it is becoming more common for data with diverse and complex formats to be collected and transmitted from numerous sources. The term Big Data is frequently used to refer to the volume, variety, velocity, and complexity [15] of data created daily. The main goal of Big Data related research and development is to manage and transform available real-time and historical data to inform decisions. The relevance of big data cuts across all sectors, including cybersecurity; indeed, it has become one of the most important considerations for organizations looking to protect themselves from cyber attacks. Therefore, organizations have made significant investments in security infrastructure and monitoring systems to protect their data and technological assets; however, due to the characteristics of big data, this is a challenging endeavor. In this context, Visual Analytics (VA) deserves special attention, as it provides advantages and opportunities for making sense of big data and informing decision making. In this paper, we describe a process to offer guidance on the design of VA tools and explain the characteristics of a tool called VKE. We reflect on its utility and potential applications.

Index Terms: Cybersecurity, visual analytics, big data, security and protection, human-computer interaction

1. INTRODUCTION

The number of conventional and unconventional cyber attacks (e.g., DDoS attacks using IoT botnets and ransomware) aimed at traditional and non-traditional targets, such as electoral processes, has increased in both number and sophistication over the past few years [3,34,43]. As a result, the need for organizations to monitor and improve the effectiveness of their security infrastructure, particularly for detecting intrusions and preventing security breaches, has also increased.

Although many efforts have been made to secure the infrastructure and data of organizations [39,40], there are still many open opportunities to improve, and thus, to contribute with research for novel methods and techniques [6]. The main challenge of organizations is to avoid attackers that compromise business data [11] and resources of IT infrastructure, which could be used as intermediary agents (zombies) to launch large scale attacks on third parties [29,48].

Some researchers claim that the main obstacles to taking full advantage of Big Data is not related to methods and technologies to transform data into knowledge, but to cybersecurity threats [11]. Thus, there is an increasing need to monitor security events for mitigating and preventing attacks [27,33]. In this context, logs from firewalls, routers, servers, computers, user authentications, network traces and events, and IPS and IDS devices are of great importance and should be analyzed to identify traffic patterns, malicious activity, and security breaches, and to contain, remediate, and recover from security attacks [8,47].

However, logs are volatile (i.e., the analysis of logs, sometimes, should be carried out within short periods of time after they are collected), large, their veracity and validity depend on the reliability of systems that collect them, and they are produced in many different formats at high velocity. Hence, such log data matches the characteristics of Big Data and in this context, analytics deserves special attention in providing insight into these large volumes of data [30], as it is located at the top of the process stack that transforms data into knowledge. According to Davenport [7], analytics has been an important player in the definition of strategic plans, and has helped to drive, for decades, the improvement of the ability of organizations to outperform competitors by means of analyzing the available data.

The analysis of cybersecurity data requires the use of advanced processing methods and techniques to aid organizations in taking action on ongoing security events, carrying out forensic analysis, and preventing future attacks [46,49]. Even with advanced data analytic methods, the results of analyzing security data can be complex and difficult to understand; thus, results alone do not provide sufficient information for effective decision-making. Consequently, Visual Analytics (VA) is a natural evolution of analytics that exploits new computational algorithms for data analysis and the ability to present and explore their results by means of visual representations. Furthermore, VA is now widely accepted as an essential tool in a wide range of domains such as Business Intelligence, security, marketing, life sciences, and social sciences.

Recently, research efforts have been aimed at using visual representations along with interaction techniques to gain insight from results [12,23,24] and the application of VA to analysis results [14,31] with the aim of providing better inputs to decision making. The goal of VA is to provide a layered filtering process in which data is successively transformed into information and then into knowled-
ge, with the active participation of analysts [19, 21]. Therefore, the application of VA to cybersecurity data is of great value [14, 31] to obtain knowledge from security events and perform the appropriate decisions to mitigate threats effects or avoid them.

Consequently, the aims of this paper are to describe the application of VA to cybersecurity data and to describe the design details of VKE, a VA toolset targeted at uncovering relationships between network traffic flows, threat types, and geographic locations from where traffic is originated, using the logs of IPS devices as the data source. The rest of this paper offers a background section (section 2), describes the process of applying VA to cybersecurity data 3, the design of VKE (section 4), discusses two scenarios of the use of VKE in section 5, presents some background details (see section 2), and discusses the main conclusions of the research (section 6).

2. BACKGROUND

VA is a process whose goal is to provide insight into vast amounts of data using advanced analytics and visualization. This method deals with scientific, forensic, academic, business, social networks [44], e-learning [18], ontologies [16], software systems [21] and temporal data [28], which is stored by heterogeneous data sources, such as databases, HTML, XML files, metadata and source code. It iteratively collects and preprocesses data, carries out statistical analysis, performs knowledge representation, provides visual representations and requires the cognition [10], perception and capacities of humans for interaction, exploration and decision-making [25, 26, 32, 42] to be successful.

VA can combine different visualization types such as hyperbolic trees, graphs, treemaps, radial, parallel coordinates and grids, to name but a few, and exploits the advantages that each one has to offer [36] to provide analysts with different levels of detail. The combined use of visualizations permits analysts to understand the relationships among elements located in separate, but linked, visual representations. In addition, Additionally, the use of multiple visualizations allow them to explore data from many different viewpoints and have more interaction paths available that may lead to knowledge discovery.

Regardless of the complexity of the problem at hand, the success of any VA solution depends on the appropriate design of the visual representations and use of interaction techniques. In synthesis, VA consolidates the advantages of machines with the strength of humans and therefore, the human is at its heart [9] and human computer interaction (HCI) is a key component to support knowledge discovery.

Some interesting works on the use of VA in the context of cybersecurity have been published recently. The authors of one of these research pieces explain that security staff and managers frequently are unable to apply security patches and fixes because they require to produce an unacceptable interruption of services if a system is restarted [1] after performing such task. Hence, personnel needs to understand the security problems to prioritize their work and select the appropriate fixes. Consequently, the researchers propose a VA tool named VULNUS to facilitate the comprehension of the network status, the impact of the vulnerabilities, and create a strategy to block the attack paths. The solution uses data collected from vulnerabilities and network scanners to trace the threats and combines several visualizations to allow the selection, inspection, filtering of vulnerabilities, and the simulation of the effect of applying different fixes to the system.

The usefulness of VA resides in the combination of advanced data analysis and visualization, which is an advantage over other analytic methods. Research has shown that the sole use of machine learning to analyze security threats is not enough as it requires the use of labeled datasets, the update of models and limitations to detect novel attacks. However, Situ is a VA proposal that uses an unsupervised machine learning method and information visualization to detect anomalies, aid the identification of attacks, suspicious behaviors, and help to understand their context [22]. This tool uses several linked views to display the analysis results and assist in the discovery of attack patterns that emerges from network flows, firewall logs, and proxy logs.

Situ is similar to VKE as both shows the flows and patterns of possible attacks using network flows and firewall logs. However, the visual representations of VKE also show type of malicious traffic, the origin of the traffic and the complete traffic flow.

3. VISUAL ANALYTICS AND CYBERSECURITY

The function of VA in the analysis of cybersecurity derived data is to move analysts towards the construction of useful knowledge through an interactive process, in which users are engaged into a productive and iterative exploration process that allows them to request additional data to the system as they identify interesting patterns. This exploration involves browsing, filtering and exploring different perspectives of data in one or more visualizations until they obtain the necessary knowledge or consider that is impossible to reach a determinate conclusion, using the data and representations available. Therefore, VA facilitates the analysis of historical and real-time cybersecurity data, for example, related to logs, IDS and IPS signatures, email data, and malicious code. The tools for these types of data include the use of representations as diverse as graphs, heatmaps, dispersion graphs, radial layouts, treemaps, parallel coordinates [17], maps, bar charts, and grids [41].

The application of VA to cybersecurity data is a transformation process that could be thought of as a funnel, where raw data is analyzed and filtered in several steps until this is converted into knowledge. The output of the process is a reduction, in terms of volume, of the original input, which contains all the required elements to carry out informed decision-making. This transformation is described in Figure 1 and is based on the VA process proposed by Gonzalez-Torres [19–21], which took as a basis the previous definitions formulated by other authors [4, 5, 42, 45].

The process mentioned above involves reading data from heterogeneous data sources, then clean and correlate data, carry out the automatic analysis of historical and real-time data, create visualizations and study the relationships between the elements involved in the results. The phases that constitute this process are Extract, Transform and Load (ETL), Advanced Data Analysis and Visual Knowledge Explorer.

ETL has the function of performing the connection and retrieval of logs from data centers, file servers, DBMS servers, web hosting servers, cloud hosting infrastructure, workstations, routers, firewalls, IDS devices and switches, and also collects data from network traffic and IoT networks operated by organizations. This phase can take advantage of software tools such as Apache Kafka, Apache Storm, Apache Pig, and Apache Tez. Apache Kafka works as a subscriber system composed of producers and consumers. Producers publish data to “topics” (i.e., a “topic” is a category name to which consumers subscribe), which could be read, cleaned, integrated, correlated, transformed and filtered by consumers, using Apache Storm, Apache Tez and Apache Pig [13, 35].

The responsibility of the Advanced Data Analysis phase is to carry out in-memory analysis for real-time data and execute batch processing tasks for historical data. Real-time data might be processed by Apache Spark, which performs analysis tasks without having to write the information directly to the hard disk, decreasing processing times [37], whereas historical data can be handled by the latest version available of Apache Hadoop [35]. The integration of in-memory and historical analysis results is a complex task that has not yet been solved successfully, despite some efforts made in this direction [2]. In this context, the most suitable options nowadays include the use of Spark and MemSQL, which can facilitate the comparison, integration, and query of results of real-time and historical

data analysis.

Once data has been processed, the results could be refined in the Visual Knowledge Explorer stage. This phase is concerned with the definition of the visualizations designs and the human-computer interaction techniques needed to support user interaction, and thus, it is responsible for creating the data models, data structures and visual mappings required to coordinate the data to be displayed.

The function of the Visual Knowledge Explorer is to move users towards the construction of useful knowledge. This stage makes possible a feedback loop between the user and the system: the user explores the visualizations, requests additional data to the system using the available interaction possibilities, and the system provides the required data according to the availability of the proper data models, structures and visual mappings. The user continues to interact with the system, browsing, filtering and exploring different perspectives of data in one or more visualizations until she obtains the necessary knowledge or considers that is impossible to reach a determinate conclusion, using the data and representations available.

There is a large number of database options for Big Data, including SQL, Non-SQL, NewSQL, and SQL-On-Hadoop [38]. So, it is essential to study in detail the particular problem at hand, according to the type of data (structured, unstructured, real-time or historical data), and the features of the available databases for storing and processing large volumes of data, to decide which option to choose.

4. VKE

A common task in security analysis is to characterize the composition of network traffic when unusually high traffic rates are coming into the IT infrastructure. The high traffic may be a result of an active attack or maybe a peak in transactions from business branches or partners. The first step in determining the origin and destination of network traffic is to use IP addresses and a map to plot their geographic locations and obtain this information intuitively and quickly, at first glance. If the traffic does not appear to come from a reputable origin, the protocols used (i.e., TCP, UDP, RTP, ICMP or other) and target ports should be determined, as well as how traffic has been classified and its low-level composition. Furthermore, if the traffic is due to email, the content of attachments should be analyzed to detect malicious code, study execution traces, and identify potential system damages using specialized tools (e.g., Cuckoo).

Inspired by the aforementioned needs of security analysis, and the benefits of VA, VKE makes use of linked views (see Figure 2), various interaction techniques, a color code, a parallel node-link (PNL) (see Figure 2 (a)) visualization for characterizing the flows of traffic details (i.e., protocols, IP addresses, ports, and traffic classification), a map to depict the location of the sources and destinations of traffic (see Figure 2 (b)), and a tag cloud to show statistics on the frequency of the classification of traffic types\(^1\) (see Figure 2 (a)).

The data displayed by the visualizations is configurable, so users can select the variables to be used by dragging them from a list of available fields to a list of chosen fields, as shown in Figures 3 (a) and 3 (b). Figure 3 (a) illustrates a user dragging a variable from the Fields Available list to the Fields Selection list, whereas 3 (b) display both lists after more variables have been moved from one list to the other. After that, the PNL visualization is rendered and the selected variables are associated with the map and the tag cloud. Figure 4 (a) shows the selection of the variable to be displayed by the tag cloud, while figures 4 (b) and 4 (c) depicts how a user selects the variables to be associated to source and target pins in the map.

The PNL visualization characterizes traffic flows using all the variables in the list of selected fields and permits analysts to get an insight of the correlation of variables and to have a clearer perspective on how traffic and data are flowing in the network. It depicts the network traffic flows with lines connecting colored squares, which are placed in columns (see Figure 5). Columns represent data categories, and colored squares depict variables within a category. The number of columns depends upon the number of variables selected by users, as described before, whereas the height of colored squares encodes the aggregated value of the corresponding variable. The color of the squares is associated uniquely to a variable. The order in which variables are added to the list of selected fields determines

\(^1\)A demo video is available at http://gonzalez.cs/flows-agtorres.avi.
the order of columns in the graphic.

The configuration of PNL in Figure 5 shows six category columns: Protocol, Src_IP, Src_Port, Dst_IP, Dst_Port, and Type. The interaction in this visualization starts when users click on a colored square and the tool highlights in orange the lines that depict the traffic flows related to the variable represented by the square. Then, users can move the mouse over a colored square, and the relevant traffic flows are emphasized using a red line. Furthermore, users can obtain, via a tooltip, the name of the variable associated with the square and the percentage of traffic associated with it. Interaction with the tag cloud and the map can provide further details of the composition of the flows, as these represent the geographic locations and types of traffic cursed.

The tag cloud complements the PNL visualization providing statistical information on the type of suspected threats found in communications. The main visual elements of this visualization are the size and color of words (the default color is red), which encode the frequency of threats classified in the Type category (or the category that was chosen to be displayed during the configuration phase). Thus, the higher the frequency of a word in the log data, the greater is its size and the darker is its hue in the tag cloud.

The PNL visualization is complemented by a map (see Figure 6), in which colored pins depict the sources and destinations of communications: red pins are used to show suspected attackers, green pins for targets, and dark blue pins for locations that could be either a source or target of attacks. The color of the pins changes according to the interactions performed with the PNL and the tag cloud. When selections in these visualizations are made, the colored point in the center of pins is also changed: a red point is used to depict suspected attackers, a green point for targets, and a gray star for those locations that could be either a source or target of attacks.

VKE was developed using JavaScript, MongoDB, Express, Angular, Node.js and Google Maps, and uses log data taken from a Snort NIPS.

5. USE CASE SCENARIOS

The interaction trajectory taken by users during analysis could start in any of the three visualizations. Thus, several different scenarios are possible. The following scenarios are some practical examples in which VKE could be used.

Scenario 1: A user wants to explore TCP traffic flows and decode source IP addresses and target ports involved in communications. The interaction between the user and VKE could proceed as follows:
Scenario 2: A security analyst is examining the recent traffic activity with VKE, and it comes to her attention that the tag cloud has registered many Trojan and malware events; thus, she decides to investigate the traffic flows and the location of sources and destinations of traffic. Accordingly, the steps involved in this investigation are as follows:

1. The user clicks over the word TROJAN, which is featured in yellow (see Figure 8).
2. The color of traffic flow lines in PNL and the pins in the map are changed to yellow, as is illustrated in Figures 9 (a) and (b).
3. The user examines PNL and determines that most traffic has been originated by 202.45.65.18 which is her own IP address (represented by the light green square) and sent over to 4 destinations. Furthermore, the IP address has also received many incoming communications, according to the size of the light green square in the Dst_IP category.
4. According to the map, traffic coming into 202.45.65.18 originated in Europe (see the yellow pin in the left side of Figure 9 (b), beside the dark orange pin), whereas the outgoing traffic was sent to two different locations in the U.S. Sources and destinations are identified by a small colored point in the center of the pins, as explained above.
5. Subsequently, the user clicks on the word MALWARE, which is painted in dark orange color in the tag cloud, whereas the corresponding traffic flow is rendered using the same color in PNL and the sources and destination of traffic are represented by pins of the same color in the map.
6. The dark orange line was drawn above a yellow line, indicating that the same traffic parties were sending and receiving Trojan attacks. However, this line brings out a comparison of the magnitude of traffic flows involved with Trojan and Malware communications.
7. The dark orange pins in the map contain a gray star in the center, indicating that these locations were both sources and destinations of malware traffic.

Based on the information provided by VKE, the user can conclude that she is a source and destination of traffic carrying out Trojans and malware. This may indicate that her infrastructure was taken over by attackers to attack third-party networks. However, an outsider analyst could also conclude that, because of the geographic location of the IP address 202.45.65.18, it is an attacker.

6. CONCLUSIONS

The analysis of cybersecurity data is one of the most important tasks that any IT department can perform. This requires the use of appropriate methods and techniques to identify threats, reduce vulnerabilities and incidents, mitigate the impact of attacks, and take remediation actions. Because the relevant data is so difficult to analyze, it is important to take into account the advantages of using VA to transform data into knowledge through advanced methods and techniques with the active participation of users.

The main contribution of this paper is the design of VKE, a toolkit to analyze security logs with the use of three linked visualizations that permit users to study traffic flows, the geographical origin of attacks, and the classification of attacks. The interaction possibilities of the tool allow users to navigate through the visualizations in an intuitive and easy way and obtain useful knowledge in a timely manner.

Future research includes the full testing of the tool with many different datasets and its validation with real usage scenarios as well as case and user studies.
Figura 7: Analysis of traffic flows and locations

Figura 8: Frequency of type of security events detected

ACKNOWLEDGMENTS

The authors wish to thank ULACIT (Costa Rica) for its support to this research and also to Juan Alvarez Piedra, Osael Josue Jimenez Murillo y Pablo Andrés Rodríguez Blanco for their contribution with code and ideas.

REFERENCIAS


Figura 9: Frequency of type of security events detected