Towards Visual Analytics for Web Security Data

Victor Le Pochat Tom Van Goethem Wouter Joosen
imec-DistriNet, KU Leuven, 3001 Leuven, Belgium
firstname.lastname@cs.kuleuven.be

1. INTRODUCTION

The continuing expansion in number and impact of cyber attacks, data breaches and other forms of cybercrime reminds us how fragile security on the Internet remains. The web ecosystem, in which these incidents occur, therefore warrants analysis to obtain an overview of the illicit operations and actors involved, ultimately in order to create better defenses.

For this purpose, security analysts collect vast amounts of data on the systems and networks they oversee. In addition, researchers set up large-scale measurements to investigate various security issues, e.g. across the IPv4 address space [2]. Finally, organizations such as CERTs and security companies monitor the cyber world and publicly share their data.

Combining all of these sources provides a wealth of data to explore, but without proper analysis tools, the task of extracting insights proves to be insurmountable. The field of visual analytics integrates visualization and interaction into the analysis process [9], using the former to leverage the increased data processing power of the human perception [6] and the latter to encourage data exploration. These methods speed up and improve data analysis and allow to handle the large quantities of data. Unfortunately, cyber analysts who feel unfamiliar with using visualizations may be reluctant to deploy them in their work environment [3]. The benefits that visualization and interactive exploration are known to have for understanding cyber security data [5] are then lost.

We propose the design of a visualization application that addresses four challenges we identified in the little explored field of web security visualization. These challenges consider the role and expertise of cyber analysts as well as the large scale and diversity of available data. We strive to create a tool based on this design that analysts will want to use to enhance their insight gathering through visual exploration.

2. CHALLENGES

Even though visual exploration would be advantageous for extracting insights from large-scale web security data, the state of the art (comprising research tools for cyber security visualization [4] and general-purpose tools such as Tableau[1]) does not provide an adequate solution. We have extracted four challenges that remain to be addressed, based on the shortcomings in current tools, the seminal work on visual analytics [9] and a user study on security analysts [1].

C1: Separation of data and visualization. Cyber analysts may find the visualization process difficult and labor-intensive [3], but general-purpose tools leave this task up to them. In addition, data is often spread out across multiple sources, such as logs, crawled web pages and public data sets, but existing tools usually do not support their concurrent analysis. Separating the data and visualization concerns reduces the analyst's effort and hides the data heterogeneity.

C2: Scalability. Data collection processes in web security analyses easily generate vast quantities of data. Directly working with the unprocessed data is infeasible: the data retrieval and transfer would be too slow, the visualization client would have to process too much data and the visualization itself would risk overwhelming the analyst or could have occluding objects [7]. However, general-purpose tools are prone to loading and displaying entire data sets, especially when interactive exploration is desired.

C3: Exploration. To be able to fully grasp the large quantities of data while avoiding information overload [7], the cyber analyst should be able to easily explore that data both on a high level and in depth. However, visualizing information well is not a trivial task, and poorly constructed visual representations could obscure interesting patterns or even cause incorrect conclusions [10]. General-purpose tools expect the analyst to select and design the visualization, even though they may lack the necessary expertise.

C4: Web security data. The relevance and interpretation of the visual representations is influenced by the characteristics of the data itself. Entities in web security data usually are of specific types, e.g. IP addresses, and follow specific structures, e.g. grouping subdomains on their second-level domain name. However, general-purpose tools are not aware of them. Meanwhile, the research tools do not cover web security, but focus on event handling in network security [4] and malware analysis [11].

3. DESIGN

The design of our visualization process focuses on tackling the four identified challenges. By integrating solutions to each one, this design will fulfill the analysis and visualization needs of cyber analysts and therefore aid them in gathering insight from the large data sets available to them.

3.1 Data abstraction

We abstract the heterogeneity of the data from the visualization (C1) through a transformation into standardized *features*. We structure a feature as a set of attributes and a data context consisting of data type annotations and human-readable descriptions of the attributes. We particularly in-

[1]https://www.tableau.com/

clude data types relevant to web security (C4), such as IP addresses for network-related analyses or domain names for cybercrime investigations. Finally, a feature contains code describing how raw data is dynamically accessed and transformed into records of the feature's attributes. This standardization allows for specialized and heterogeneous sources, especially those commonly used in security analyses such as log files, to be easily integrated with our visualization tool.

3.2 Aggregation

Through aggregation, we improve performance and scalability (C2) by reducing data size and processing and therefore response latency, as well as increase interpretability (C3) by reducing the risk of overloaded visualizations. Data is aggregated during transformation, e.g. by grouping similar items on an attribute value or with clustering algorithms, which can exploit structures within the data to minimize information loss. Web security data often contains such hierarchies (C4), e.g. autonomous systems for IP ranges. Data processing and transfer are minimized by aggregating data early on during the retrieval phase and only retrieving detailed data lazily upon explicit selection, as opposed to retrieving all detailed data upfront and loading it in memory before aggregation.

3.3 Interactive visualization

Visualization presents more data in a single view while maintaining understandability (C3). To abstract the visualization concern away from the analyst and reduce their effort (C1), we automatically derive the most appropriate representation from the data types. To allow correct and easy interpretation, our charts follow best practices from information visualization [10]. Specialized visual representations allow to study patterns specific to web security data (C4).

Interaction allows the analyst to manipulate what and how data is being shown, highlight interesting areas and study them in more detail. To reduce data overload, a visualization initially shows an aggregated overview of the data. The analyst can then interactively zoom into and select a subset of the data, to obtain more detailed data on demand [8].

We arrange multiple visualizations for separate features on a dashboard. We implement the synchronization of interactive selections ('brushing') for mutual refinement ('linking') [12] across these charts. We also add interactive joins of two features, to enhance the exploration of correlations.

3.4 Integration with public data

The analyst can enrich their data with public security data (C4). We provide interactive retrieval of such data, and allow exploration through a transformation into features.

Public data sources differ in how data is retrieved: either the whole data set is downloaded (e.g. the Alexa top 1 million sites[2]), or the data is obtained through an API (e.g. VirusTotal[3]) and therefore usually subject to rate limits. We allow for queries and aggregates for API data sets by incrementally acquiring data. This also applies to web crawlers, which can be seen as equivalent as data is acquired one page at a time.

In addition to manual requests to the API for individual items, we support background data preloading. The analyst interactively queues items based on selections in another data set. A worker then processes this queue, requesting and storing the API's data for each item. The worker can space out requests in time to take into account any rate limits. We support visualizing partially acquired data, and by intelligently traversing the preloading queue this data can be representative for the complete data set. This means preliminary conclusions can be drawn much quicker, potentially avoiding the need to load the full data set.

[2]https://s3.amazonaws.com/alexa-static/top-1m.csv.zip
[3]https://www.virustotal.com/

4. IMPLEMENTATION

We have implemented a prototype of our design, whose architecture follows the client-server model: the server handles the data retrieval, while the client handles the visualizations and interactions used for exploring the data. The data storage itself is handled separately and can be tailored to the characteristics of the data. This model places the burden of retrieval and processing of the raw large-scale data on the server, which reduces the processing power needed on the client. The visualization client is web-based, allowing analysts to explore data across devices and platforms.

5. CONCLUSION AND FUTURE WORK

We introduce the design of a visualization application that tackles the challenges that arise when exploiting the benefits of interactive visual exploration for analyzing the abundance of web security data. Our future work consists of continuing the development of our prototype by refining and expanding its functionality. We seek to improve data access across technologies, simplify feature definition, integrate more visualization types and add interactive data analytics.

In the long term, we plan to make our visualization tool available to other researchers and analysts, as a platform for encouraging collaboration on data analysis. This opens up more possibilities to analyze ecosystems, test hypotheses and gather valuable insights from the wealth of available data.

Acknowledgments. This research is partially funded by the Research Fund KU Leuven.

6. REFERENCES

[1] D. M. Best, A. Endert, and D. Kidwell. 7 key challenges for visualization in cyber network defense. In *Proc. VizSec*, pages 33–40. ACM, 2014.

[2] Z. Durumeric, D. Adrian, A. Mirian, M. Bailey, and J. A. Halderman. A search engine backed by internet-wide scanning. In *Proc. CCS*, pages 542–553. ACM, 2015.

[3] G. A. Fink, C. L. North, A. Endert, and S. Rose. Visualizing cyber security: Usable workspaces. In *Proc. VizSec*, pages 45–56. IEEE, 2009.

[4] F. Fischer. *Visual Analytics for Situational Awareness in Cyber Security*. PhD thesis, Univ. Konstanz, 2016.

[5] J. R. Goodall. Visualization is better! A comparative evaluation. In *Proc. VizSec*, pages 57–68. IEEE, 2009.

[6] D. A. Keim. Visual exploration of large data sets. *Commun. ACM*, 44(8):38–44, Aug. 2001.

[7] D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler. Challenges in visual data analysis. In *Proc. IV*, pages 9–16. IEEE, 2006.

[8] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proc. VL*, pages 336–343. IEEE, 1996.

[9] J. J. Thomas and K. A. Cook, editors. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE Computer Society Press, 2005.

[10] E. R. Tufte. *The visual display of quantitative information*. Graphics Press, 1983.

[11] M. Wagner, F. Fischer, R. Luh, A. Haberson, A. Rind, D. A. Keim, and W. Aigner. A survey of visualization systems for malware analysis. In *Proc. EuroVis - STARs*, pages 105–125. Eurographics Assoc., 2015.

[12] M. O. Ward. Linking and brushing. In *Encyclopedia of Database Systems*, pages 1623–1626. Springer, 2009.