Visualization of Malware across Business Sectors

Emily Longman, Will Sims, and Taylor Kirkpatrick

Abstract— Businesses are targeted by cyber attacks every day, and there is no efficient way to visualize the industries that are most at risk and the types of attacks they are vulnerable to. Our research focuses on constructing a useful visualization for security professionals so they know what industries are most in need of their services. Currently, there are no interactive visualizations that solve this specific security problem. However, we built on previous research that looked at different problems in security visualization. The benefits of this visualization is to provide a tool that allows users to easily discover sectors most at risk so we can work towards keeping peoples information safe.

Index Terms—information visualization, security, malware, phishing

1 Introduction

Cybersecurity is a serious concern, or at least should be, for almost every industry. In the modern world relies heavily on computers of all forms, for basic tasks like checking the weather to heavy industrial computation. The data stored and processed by these systems can be very valuable and targeted for theft. Because of this there has been an ever-growing list of malware used to break into computer systems and gain whatever the attacker was looking for. Today, with nearly 40 years of enterprise computer usage, there is a lot of data available on the exploits, but it can be very technical and hard to sift through. With information visualization we can do a better job on analyzing these trends, which allows the creation of better defenses against future attacks.

2 Related Work

We looked at research in the field of cybersecurity and information visualization to see any problems people encountered and how we can avoid those challenges when creating our own visualization. We were also interested in seeing if there has been any visualizations on the same topic and if we could build on that work for our own project.

2.1 Challenges in Visualization for Cyber Security

Gates and Engle looked at why cyber security visualization has not been more effective in the past, how visualization can be utilized in cyber security, and how to evaluate cyber security visualization [4]. One of their main points is that you should not create a visualization for the sake of visualization and you should always have a clear problem that you would like to solve. Aesthetics should always come second to displaying information clearly and concisely. They also emphasize how the process should always be evaluated and case studies are sufficient for valuable feedback instead of large user studies. Best, Endert and Kidwell looked at seven key challenges when creating visualizations for cyber security [1]. The challenges that they encountered when developing a network security visualization that are most related to our problem had to do with their data sources. One of the issues in cyber security visualization is that there is so much data that only a small subset of data is compared when analyzing an event. This can be problematic because important information can be left out. Additionally, The diversity and quantity of security information provides challenges and it is important that users take into account the scope and validity of their data.

2.2 Evaluation Methods for Cyber Security Visualizations

It is critical to evaluate visualizations in order to determine whether or not the product is usable and solves the problem that is addressed. Langton and Baker explored various evaluation methods specifically in the domain of security and provided recommendations based on their test results [5]. It was found that design heuristics and user studies are effective for creating information visualizations in the field of cyber security. The methodology discussed for measuring the complexity of computer security visualization designs was developed by Suo, Zhu, and Owen [11] and they tested this method by evaluating TNV and RUMINT cyber visualization tools. Metrics such as color, shape and size were used to evaluate the complexity of the visualization. The user testing consisted of using common metrics such as task success, time on task, and errors. [9]

2.3 Design in Information Visualization

According to Moere and Purchase, good design is a fine balance between utility, soundness, and attractiveness [13]. The requirement of attractiveness was investigated in and a model of the potential roles of design in information visualization was created. The model is made up of three distinct categories of visualization: practice, studies, and exploration. This model was based off the interaction design triangle developed by Fallman [3]. It was also recommended that pastel colors are used in visualizations to add less eye strain on the user. The journal article Fluid interaction for information visualization explored different types of interactions to consider when designing a visualization and created a set of guidelines for creating fluid interactions [2]. Some of the main guidelines consisted of using providing immediate visual feedback on interaction, minimize indirection in the interface, integrate user interface components in the visual representation. The guidelines closely followed Nielsens 10 usability heuristics [8]. Lastly, The design of our visualization was influenced by the stock market visualizations that organized all of the data by industry sector [10].

3 Method

A variety of methods went into the creation of this visualization, ranging from color theory to empirical research. Much of the visualization ideas were taken from preexisting security data compilations. To define the best way of displaying this dataset, much analysis was done to determine which similar ones were most effective. This was done by assessing which infographics were easiest for users to understand.

3.1 Method Overview

To assess the preexisting visualizations, users unaffiliated with the subject and with security were asked to answer question using them. The ones with which the users were able to most easily answer questions were the design ideas that were implemented. The aspects which did the best and were implemented are as follows:

- Color usages associated with positive or negative traits (specifically red and green)

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

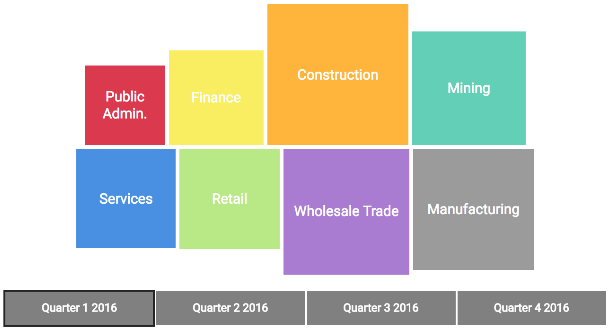

Fig. 1. Final design for the layout of the different industries and number of attacks.

- Visually simple data
- Multiple levels of increasingly specific information
- Infographics rather than graphs

All of these were combined in some way to create the final version of the visualization. The same tactics that were used to assess the preexisting security visualizations were also used for the assessment of the final product, as is described in the Results and Performance below.

3.2 Visualization Tasks

The question we are trying to answer is what industries are most vulnerable to cyber attacks and what types of attacks are most prevalent. This problem is important because it provides a way for cybersecurity companies to see which industries are most in need of their services. It also allows industries that are most affected by cyber attacks to see which types of threats they are most vulnerable to.

3.3 Data Sources

There are plenty of security companies that disclose information about cyber attacks. Popular cyber security companies such as Symantec, McAfee, Malwarebytes publish monthly and yearly threat reports about different kinds of attacks. Symantec organizes the monthly threat report into five different categories: Malware, Web Attacks, Spam, Phishing, Mobile, and Social Media. They also give information about the number of threats per industry [12]. Malwarebytes publishes a state of malware report every year which contains information on the most common types of attacks with a focus on ransomware trends in particular. They give trends of different threats over the last three years, which provides a more thorough data set. [6] In the McAfee report, the volume of malicious samples cataloged per quarter ebbs and flows quarterly and annually. This data showed a decline during the past three quarters which mirrored the trend we observed at the start of 2015. A pattern of two to three quarters of growth followed by three quarters of decline has been consistent since 2013. [7] For the purposes of our project, we used the data from Symantec for the visualization.

3.4 Design Comparison

There were many iterations of design ideas that went into this visualization page, ranging from grand scheme design choices to small implementation tweaks. This basic framework of having boxes represent different industries stayed throughout the process. Each of the industries has their own section, and these all contain data about which types of attacks are most prevalent. We had initially planned on lining up the boxes from biggest to smallest, but decided to use a cluster of the boxes instead so you can view more information at once. This is also more aesthtically pleasing. Our final design to represent the number of attacks for the industries in Figure 1 took inspiration from the stock market visualization in Figure 2.

Fig. 2. Our design took inspiration from this visualization on stock market data by industry.

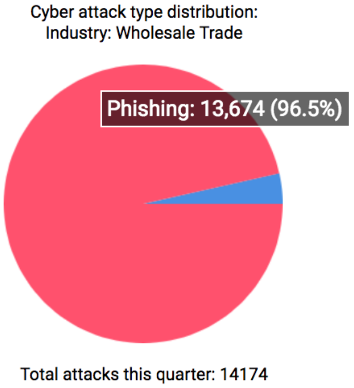

Fig. 3. Final design for the breakdown of the types of cyber attacks within an industy.

We came up with a few different ideas for how we would show the distribution of cyber attack types in each industry. We considered putting pie charts of the data inside of the boxes, but felt that this was confusing and it would be unclear what the the box represents. This problem was solved by placing the pie charts below the industry boxes, and you can view the attack type breakdown by clicking on the box that corresponds to the industry which you can see in Figure 3.

Initially, we had planned on putting the total number of attacks inside the boxes, but some of the boxes ended up getting so small that we didn't want to clutter the interface. We settled on putting the total number of attacks below the pie chart. The sizing of the boxes already allows users to get an idea of how often an industry is targeted by cyber attacks. We also considered showing data for every month of the year, but felt quarters would be more useful since it allows you to see trends over three months.

4 Enhancements

There were various improvements which were implemented that was different from out previous design. Instead of using a scale to color all the boxes based on size, we decided to use unique pastel colors for each box to show they are different industries. Past research has suggested that pastel colors are effective and a color gradient didn't make sense in this context [3]. Another enhancement was adding borders to the active industry and quarter in the visualization. This improvement provides feedback to the user about which component of the interface is active. Feedback is an important design heuristic based on the research done by Nielsen [8]. One improvement was sizing the boxes relative to the total number of attacks instead of static numbers. By using dynamic sizes that change based on the number of attacks each quarter, it allows users to accurately compare industries and look at the variability.

5 Design Comparison

There were many iterations of design ideas that went into this visualization page, ranging from grand scheme design choices to small implementation tweaks. This basic framework stayed throughout the process, but the design of it evolved over time. A scroll bar for different months of data was one of the big things that was added, as well as a pie chart breaking down the data within each category upon clicking on one. Each of the industries has their own section, and these all contain data about which types of attacks are most prevalent.

A number of arrangement ideas for the data were compared, with the look eventually settling on one of differently sized and colored blocks. These help to display the distribution difference both with color and size, so that someone with some form of visual impairment is likely to still be able to tell the difference. There was much discussion over the use of a red to green scale, since a number of colorblind people may be affected, but it made the most sense for our data, and the size difference help to provide a backup for this.

6 Implementation

Our original plan for implementation before we started was a standard front end with a mySQL back end. The database would contain data on the number and type of attacks per industry, per month for the year of 2016. We quickly found that this was not feasible, as we did not have the resources or knowledge necessary to make a persistent database for the project, and the data was of such scale that entering it manually was too big a task. The data was not in a friendly format for scraping, and so we had to make our scope easier to handle. We decided to instead use a JSON format for the data, and organized the data by quarter instead of by month. The JSON data (contained in the scripts.js file) is much more manageable and simple to use, getting the job done as well as a database would without significant overhead. Organizing the data by quarter also may have improved the visualization, as cluttering the screen with twelve options to select may be more confusing for the user.

As development was underway, the design changed greatly once we could actually test things interactively, and we found limitations and non-obvious problems with our design. One of the earliest issues noted is, after proposing a horizontal scrollbar to change between months, users seeing the visualization on a smaller screen would have multiple scrollbars on the screen. The point of the visualization is to simplify information presentation, not make it more complex, and so we had to change this. Our initial design featured a single line of boxes that functioned similar to how they do now, but this was the root of the issue as the constraint to a single forces an odd looking view for all users without sufficiently wide browsers. In addition, while observing the design, we found that it might be confusing for the industry boxes to move around as much as we proposed they would, as users may not be aware the boxes no longer were for the same industry. We remedied this with a significant change in design, by forcing a second line for half the boxes. This change was one of the largest deviations from the design. This change later necessitated another change to the design to make it more fluid and visually appealing. Instead of having the boxes in both rows simply grow and shrink in place, we made all of the boxes aligned near the center of the figure, and never changing position. While not only visually appealing, this better allows something we were missing in previous designs, the highlight of the boxes relative to each other. With boxes that are changing position it is difficult to compare them to other boxes, and do an industry to industry comparison. When the boxes stay in place, however, it is easy to compare them to each other. They also now slightly shift position in response to the change in size of their neighbors, keeping the figure looking fluid and dynamic. Because of the shape of this new figure, we had to change how the popup menu works as well.

The pie popup in the design was changed in favor of a static location on the page. This was done to avoid covering up any boxes, as such a thing was a possibility with the new changes to the figure layout. Moving the pie chart to a static location also allowed us to add extra information that was lost when we removed the third menu in our second design. In addition to being a cleaner and more informative method, it also allowed us to use the jQuery plugin Sparklines by Gareth Watts to generate the pie chart easily. It would not be as easy if the pie chart had a dynamic location. As for the pie chart itself, rather than use a key, we took advantage of the package and used the hover labels included. The colors used in the pie chart are colors taken from two of the boxes. These box colors are pastels, which promote attention to the visualization and are also good colors that are not jarring, yet still distinct enough to avoid confusing the boxes. The original design included colors as a heat map, but it was found that unless we wanted to set arbitrary number to change color at, heatmaps were unnecessary as the information is already presented in number and in the size of the boxes.

7 Results and Performance

After having fully implemented the visualization it was tested on a number of users unfamiliar with both security and information visualization. These users helped to provide useful feedback on both the clarity and effectiveness of the final product. To do this test users were asked to answer a few simple questions with the data interface provided. For instance, they were asked ”Which industry had the most attacks in the spring of 2016” which they would need to interact with the visualization to answer. Their usage was observed while completing these tasks, which allowed for assessment of which portions might be confusing or misleading.

8 Conclusions and Future Work

The visualization was effective at showing the industries that are most targeted by cyber attacks. The size of the cell was used to show how many times a certain industry is attacked in a given financial quarter and users were able to easily understand. Future work may include exploring what other types of information we can put in the boxes such as the change since last quarter, total attacks, and types of attacks. Due to time constraints, we were unable to test these ideas in our design. This topic has a wide variety of data sources which can be combined and compared in interesting ways to draw a variety of conclusions. With database manipulations and interesting front-end visualization this raw data can be made into something very interesting. The better the visualization, the better users can understand the information and gain valuable security knowledge. The goal of this project is to provide a tool for security professionals so they can better understand which industries need improved security measures.

References

[1] D. Best, A. Endert, and D. Kidwell. 7 key challenges for visualization in cyber network defense, 2014.

[2] N. Elmqvist, A. Vande Moere, H.-C. Jetter, D. Cernea, H. Reiterer, and T. Jankun-Kelly. Fluid interaction for information visualization, 2011.

[3] D. Fallman. The interaction design research triangle of design practice, design studies, and design exploration, 2008.

[4] C. Gates and S. Engle. Reflecting on visualization for cyber security, 2013.

[5] J. Langton and A. Baker. Information visualization metrics and methods for cyber security evaluation, 2013.

[6] Malwarebytes. State of malware 2017.

[7] McAfee. Threat reports.

[8] J. Nielsen and R. Molich. Heuristic evaluation of user interfaces, 1990.

[9] T. OConnell and Y.-Y. Choong. Metrics for measuring human interaction with interactive visualizations for information analysis, 2008.

[10] B. Shoemate. Stock market visualizations, 2015.

[11] X. Suo, Y. Zhu, and G. Owen. Measuring the complexity of computer security visualization designs, 2007.

[12] Symantec. Symantec monthly report.

[13] A. Vande Moere and H. Purchase. On the role of design in information visualization, 2011.