Cyber Security Project

MIS 510

So Yeon Kim
Ryan Chinn
Gyudong Cho
Sagar Samtani
Section 1: Introduction and Background

Cyber security is a growing area of concern for both individuals and companies alike. It is for that reason that cyber security research is absolutely critical. It is important to ask the right kinds of questions and attain results which will allow individuals to gain a better understanding of potential threats and create actionable plans around them. **The purpose of this paper is to provide the reader with four different research questions and the appropriate methodologies to determine meaningful and insightful results for these questions.** In order to develop a variety of topics the four questions will be split evenly across two different data sets of Hacker Web and Shodan.

Hacker Web is a MySQL database which contains 18 different hacker related forums. Each of these forums has four components to it: authors, threads, posts, and code. Authors create posts which are stored in threads. A single author can make many posts, and many posts are part of a single thread (with a thread containing multiple posts). As part of the posts, an author can also attach code to these posts. It is important to note (and will be delved into later) that five of the most highly populated tables in the database contain about 75% of all of the data in all of the forums combined.

Shodan is a search engine for the Internet of Things (IoT). It has the capability to search for many devices which are connected to the Internet. This can be based off of the software, hardware, IP address etc. With a powerful tool like this, many exposed devices can pose significant threats. However, it is important to note that Shodan has its limitations, with one of them being its inability to detect devices behind NAT firewalls. For example, if one is trying to determine all the devices of a certain operating system, their search would bring up a significantly lower number than the actual number of devices.

As previously mentioned, the exploration of both the Hacker Web and Shodan will provide the reader with meaningful and insightful results. This paper will take the reader through the questions, hypothesis, definitions, literature review, research design, and findings/ key takeaways. In addition, a set of supplemental slides is provided in order to cover some extra areas.

Section 2: Research Questions, Hypothesis and Importance

**Research Question 1 (Hacker Web):**

Are the top five posters from each of the five biggest forums the main posters of the source code related to networking related attack topics?

**Importance** - Answering this question will enable professionals to quickly triangulate the most likely threat vectors in the case of an attack (reactive measures) and/or will enable the professionals to create a list of “potential threats” (preventative measures).

**RQ1 Hypothesis:** The top five posters of each forum are the main contributors of the overall number of posts in a forum and of the source code pertaining to networking related posts.

**RQ1 Definitions:**

Top five posters - number of total posts (in terms of volume) which an author has made.

Majority/Main contributors - If the combined percentage of the top five posters for the given topic, forum, source code etc. is >50%, then the top five posters will be classified as the main contributors, as they will have posted the majority. Please note that this definition will be applied to other hypotheses as well.

Networking related posts are the keywords of: firewall, IP, port, server, password, virus, and Trojan.

Source code is embedded source code which an author can provide in addition to their posts and attachments. This is critical because if there is any dangerous code, then it has the potential to spread very quickly in a hacker community. Furthermore, source code is one field which there is no language barrier.
**Research Question 2 (Hacker Web):**
How does the behavior of posters within forums compare in the larger threads vs the forum as a whole?

**Importance** - The findings of this question will provide an avenue to understanding what motivates the hacker community in terms of their interactions of groups. By determining their interactions in the largest of groups, professionals can gain a better understanding into the psyche of the hacker community. It will also give insight to the level of contribution which is present within the community.

**RQ2 Hypothesis:** It is hypothesized that the length of content which is posted in the top thread of the biggest forums are longer than the average length of content which all of the authors post in the forum as a whole.

**RQ2 Definitions:**
Length of content- the number of characters which are in a given post
Top thread- the thread within a forum which has the most posts in it.

**Research Question 3 (Shodan):**
How does the number of publicly facing hospital devices change over time, and what is the distribution of these devices across organizations?

**Importance** - By properly evaluating the findings of this question, industry professionals can have a better idea of the current trends in the market, know which hospital systems are properly securing their devices, and make take the appropriate action in order to get the industry moving in the proper direction.

**RQ3 Hypothesis:** With the increased emphasis over the years on device security and the more stringent privacy policies in the healthcare system in the USA, it is expected that the number of publicly facing network devices has decreased over time. Furthermore, it is expected that the industry as a whole all contribute a similar amount rather than a subset contributing the majority of the open devices.

**RQ3 Definitions:**
Decreased over time- In order to be categorized as decreased over time, the number of open devices in 2014 must be less that of 2010. 2010-2014 is the defined time range.

Subset contributing majority- This is defined as <=25% of the organizations contributing >50% of all of the devices discovered.

**Research Question 4 (Shodan):**
What is the breakdown of the HTTP response codes from hospital related devices, and what does that indicate about their accessibility?

**Importance** - Even though Shodan can detect the publicly facing devices, not all of them are going to be vulnerable to attack. The answering to this question will provide insight into what the breakdown and characteristics of healthcare related devices, and whether any action should be taken.

**RQ4 Hypothesis:** It is expected that the majority of the healthcare related devices which Shodan is able to find will provide a 400 or 500 level HTTP response, thus denying any access into those devices. Furthermore, it is hypothesized that the majority of these devices will be operating on port 443, thus using Secure Sockets Layer (SSL) and providing an extra layer of security over the traditional port 80.
**RQ4 Definitions:**

400 or 500 level HTTP response- According to W3, 400 and 500 HTTP level codes are defined as client error and server error respectively [1]. Both of these will deny users access any device or webpage.

Secure Sockets Layer (SSL)- According to Digicert, SSL is “standard security technology for establishing an encrypted link between a server and a client [2].” SSL allows for secure transfer of sensitive information.

**Section 3: Literature Review**
The foundation for this research is based on the two domains:

*Hacker Communities*

Whilst the monetary incentive might be the strongest motivation for hackers to be active in hacker communities, there could be other driving forces for them to be involved in social networks such as recognition. Hsinchun Chen and Victor Benjamin maintained in “Securing Cyberspace: Identifying Key Actors in Hacker Communities” that peer approval in hackers communities is valued heavily by members due to the nature of group dynamics [3]. In addition, Eric Raymond explained Hackers’ attitude in his online article “How to become a hacker” that Hackers believe in voluntary mutual help and should act like it is their attitude in order to be accepted by peer hackers [4].

*Internet of Things (IoT)*

The Internet of Things is the notion that ordinary objects have embedded technology which allows them to sense and interact with the environment. According to Gartner, the Internet of Things will reach 26 billion devices by the year 2020 [5]. As the number of devices connected to the Internet continues to increase, security will be a growing concern. One of the biggest opportunities and threats to the IoT exists in the healthcare space as healthcare cybercrime is becoming a reality, as demonstrated by the SANS Health Care Cyberthreat Report [6].

**Section 4: Research Design**
The methodology used to answer all the questions are broken down into three levels of analysis, with increasing levels of granularity at each level. The main tools which are used for Hacker Web are MySQL, HeidiSQL and Excel. All of these tools are used for query generation and analysis. Before the research questions of Shodan, a section will be provided on how this data was gathered in an automatic fashion. However, additional tools were used to complete the analysis, including Excel and Tableau.

**Research Question 1:**

*Analysis Level 1: Determine the five largest forums and their top posting authors*

These forums must fulfill certain criteria. First and foremost, they must have all of the requisite attributes necessary in order to complete the rest of the analysis. For example, if one of the largest forums does not have some critical attributes, or has null values in critical fields, then it will be left out of the equation. It is for this reason that the forum of mihandownloads (even though it has third most rows out of all the forums) must be left out. Mihandownloads does not have any data in its code table, a critical component of this analysis.

HeidiSQL’s provision of statistics (number of rows, size) about all of the relations in the table allowed for the determination of which tables were the largest. This, in conjunction with the aforementioned requirement of attributes narrowed down the list of tables. From there, an SQL query narrowed down the top five posters for each forum. The figures below illustrate the ratios of the tables selected compared to the whole database, as well as the queries used to determine the highest ranked authors.
The 74.8% total row coverage which these tables provide ensure that there is a large sample size which can help to infer the behavior of the rest of the forums. It should also be noted that those authors who were anonymous were treated as a group and were not included in the top five authors.

Analysis Level 2: Determine the percentage of postings and source code from these authors.

After determining the largest forums and their top contributing authors, the next step is to determine the percentage of contributions of the top authors in each of the forums. A graph with the results as well as the formulas used to find the results can be found below. Please note that in order to find the number of posts for each author (as well as overall), simple SQL was used. These are SELECT count(postid) FROM <tablename> WHERE author = <choose author>. A more complex query can be developed in order to calculate the same information, but the following method produces more flexible and reliable results. It is critical to ensure that the query is applied to all applicable authors, and that the authorID is used (to minimize the amount of potential error for trying to find the appropriate info about authors.

\[
\text{Percentage of total posts} = \frac{\sum P_j + P_k - P_l + P_m + P_n}{\text{Number of total posts}} \times 100
\]

\[
\text{Percentage of total posts} = \frac{\sum C_j + C_k - C_l + C_m + C_n}{\text{Number of total code posted}} \times 100
\]

\[P_{a(x)}\] and \[C_{a(x)}\] – Both of these refer to authors, respective to posts(P) or code(C).

Analysis Level 3: Keyword search

The final level of analysis is a keyword search using the predefined set of words. This is done through SQL queries. While the text parsing with SQL queries can be tricky, it provides a level of automation and precision which cannot be rivaled by simply reading the lines of text. After determining the number of posts which have the appropriate keywords in it, the next step is to break down each of the posters percentage of total posts related to those topics. The SQL queries which were developed are shown below.

The query on the left brings up all of the posts which are related to those specific topics. The one on the right brings up all of authors and the source code which they have posted specifically to these topics. This query can be further manipulated by adding in another AND condition which can narrow down by the authorID. Please note that both of these queries are templates which can be applied to the specified tables. These FROM and WHERE clauses in each of these queries need to be manipulated in order to glean all of the pertinent information.
Research Question 2:

Analysis Level 1: Determine the top five forums

For the purposes of this level of analysis, the same methodology will be followed as the first level of analysis in for the first research question. Those same forums will be the same ones used for this question, as they all fulfill the criteria necessary in order to conduct meaningful analysis.

Analysis Level 2: Determine the largest threads (in terms of posts) and the length of their posts

A simple SQL query determining the max is necessary in order find the maximum number of posts within a thread. From that, the threadID must be used in order to query the posts table in the most optimal manner.

To determine the length of the posts within the biggest thread, the following query template is needed. ‘SELECT * FROM posts WHERE threadID = ‘threadID’ORDER BY cast(numofposts as unsigned int) desc;’ The cast function is used in because the numofposts field is stored as a varchar, not as a number. In order to conduct any sort of average function (which is done by exporting the results set into Excel) the data which is in the database must be converted into the appropriate data type.

Analysis Level 3: Determine length of average post

After getting the information about all of the posts within the top thread, the next step is to determine the length of all of the posts within a forum as a whole. In order to do this, another SQL query was developed, and the template is as follows: ‘SELECT * FROM thread ORDER BY cast(numofposts as unsigned int) desc;’ After getting this data, it is then necessary to compare it against the average length of posts within the highest thread.

Shodan Automated Data Collection

In order to automate the data collection process, the Shodan API was connected using the Python library. A script was then developed to retrieve the desired results and store them in a MySQL database. For the input to the script, the base search parameter on the IP address ranges of the largest healthcare providers according to Becker’s Hospital Review [7]. Healthcare providers of interest include: Adventist Health Systems, Avera McKennan, Banner Health, Baylor Health Care Systems, Carolinas Healthcare System, Columbia Health Care, Dignity Health Care, HCA Hospital Corporation of America, Intermountain Health Care, Mayo Foundation for Medical Education and Research, Mercy Health System, Sanford Health, South Dakota Network, Sutter Health, and Tucson Medical Center. After executing the script, the resulting test bed consists of 1,959 records with device information such as IP address, timestamp, location data, and associated ports.

Research Question 3:

Analysis Level 1: Trend determination

The first level of analysis is based around determining what the current trend of exposed devices is. In order to do this, a simple SQL query is used in order to organize all of the devices by their corresponding year. This query was ‘SELECT date_format(timestampx, ‘%Y’), count(searchID) FROM shosearch GROUP BY date_format(timestampx, ‘%Y’);’ This query brings up the different years which are contained in the data and the amount of devices which were found within that year. After determining this information, it is critical to export it into Excel and conduct simple tabular analysis.

Analysis Level 2: Distribution of devices across organizations
After finding the trend of devices across the years, a more granular approach is needed to determine the answer a piece of the initial research question. In order to get this distribution, the following query was developed. ‘SELECT org, count(searchID) FROM shosearch GROUP BY org, ORDER BY org.’ After determining the results of this query, the data again needs to be exported to Excel, where basic analytical techniques were used in order to determine the breakdown of organizations.

Analysis Level 3: Geographic visualization and distribution

The final piece of the design of this question is the appropriate geographic visualization of the data which was collected from the first level of analysis. In order to complete this, a tool called Tableau is used. Since the data which is gathered in the Shodan results includes the latitude and longitude of all of the instances of data, the Excel sheet which was created earlier is used as a data source, and the visualization tool in Tableau to chart the data points is called symbol map.

Research Question 4:

Analysis Level 1: Classification of HTTP response codes

The first step of determining the answer to the research question is to determine the different classifications of the HTTP response codes which are present within the data set. In order to do this, the data set which is present must be evaluated in order to determine the different types of codes which are present within the data. After doing so, the following template of the query is required determine the amount of times a device has that specific HTTP response occurs within the data set. ‘SELECT datax FROM shosearch WHERE datax LIKE ‘%302 Found%’;’ Please note that the 302 Found is just an example of a code which is used. A more complex SQL can be developed in order to gain the count of all of the different codes which were found.

Analysis Level 2: Testing and placing of HTTP response codes

Based on the previous section of analysis and the aforementioned resource provided by W3, some of the codes are tested in order to determine whether or not they were accessible by web browser. This could not be an automatic process, as some of the codes to be tested were based on the uncertainty of whether or not the web pages were reachable or not. Based on their reachability, each of the codes were placed into the bin of accessible or inaccessible. Basic statistical analysis allows for a great depth of information such as a breakdown and comparison between the different types of devices in terms of their reachability.

Analysis Level 3: Determination of completely secure, inaccessible devices

After determining the completely secure, inaccessible devices, the following query was developed. It should be noted that in order to answer the initial research question only the inaccessible devices were needed. As previously noted, the devices which are on port 443 are at a much higher level of security as compared to port 80.

Section 5: Findings and Discussion

Research Question 1: After conducting the three levels of analysis on the forums discussed in the research design section, it was determined that the top five authors do not provide the majority of the source code or the amount of posts within their respective forums. It was determined that the initial hypothesis was incorrect. In fact the number of posts which the top five authors made was no greater than 5.285% across any of the forums analyzed. In terms of the amount of source code which an author
posted about the aforementioned networking topics, the most a single author posted about those topics within their given forum was **25%**. Both the 5.285% and the 25% fall significantly short of being the majority as defined in the initial experiment. A more detailed breakdown can be seen below. The left chart provides a powerful understanding as to the numbers associated with each forum.

The results of this experiment provides security professionals with a better understanding of the potential threat vectors which are associated with networking related posts. When attempting to triangulate the authors who are most likely to post dangerous code, they will be able optimize their searches for preventative or reactive measures.

There were several key lessons learned throughout this process. First and foremost, when dealing with such data sets in the future, instead of taking the top five posters, a more comprehensive analysis will be conducted by taking the top five percent of posters, or a number greater. By doing so, it will allow for a higher level of precision and insight into where the main postings of source code are originating from.

The potential further analysis which could be conducted for this research question could be looking into the specific contents of the each of the different posts and source code and determining whether it can be classified as dangerous or harmless. However, as no automated measure was developed to conduct this analysis for this paper, it was left out of the analysis.

Although it was not in the scope of the initial question, further tests were run to determine whether or not some of the devices which were found in the results of Hacker Web yielded results in Shodan. After conducting some analysis, it was determined that were a few that came up. Those are listed below, in the order of: **MAC Address 00:03:ff:f6:fd:00** (left picture) and server information of domain ashiyane.org.
Research Question 2:
After conducting analysis on initial research question, it was determined that the hypothesis which was developed was incorrect. The length of the posts within the top thread in a forum is no longer than the length of the posts within the forum as a whole. This speaks to the behavior which hackers have when they are part of a thread which has a large number of posts- they are less likely to post longer content than the average in the bigger forums. This may mean that they believe they do not have as much to contribute, or do not have to contribute as much because of the amount which others are contributing. The graph above helps to visually depict the specific breakdowns within the forums and their top threads.

As far as lessons learned and potential future analysis is concerned, it would be interesting to see a more comprehensive look at all of the different forums across a variety of different topics (not just Hacker Web). It would also be interesting to look at the specific psychological factors which affect such behavior.

Research Question 3: After completing the steps in the research design, it was determined that amount of open devices is increasing as the years pass by (with the exception of 2012, when the number decreased.) A visual depiction of how this trend occurs over time can be seen on the right. Between 2010 and 2012, this trend stayed constant, but over the past year and a half, the amount has risen dramatically. While this can be attributed to the quality of the crawler which the Shodan developer has been improving on, there is no way to know for certain.

In addition to the graph, a geographical analysis was conducted and is presented on the left. The top map depicts the location of all of the devices across the country in the year 2010 (the first year of analysis). The second map depicts the number of healthcare related devices in the year 2014, along with the population increases for the country as a whole. As it can be seen, the percentage of the number of devices increased in comparison to the population, and the dispersion of the devices across the country also increased.

By answering this question, individuals will now understand the current trends which are present out in the industry. As a result of knowing these trends, healthcare
professionals, security experts and law making bodies can make the appropriate adjustments to ensure the top level of security for these types of devices. However, for the purposes of this paper, the hypothesis which was initially developed is incorrect.

Research Question 4: After conducting research, it was determined that the initial hypothesis was incorrect. The total number of devices which were running on port 443 was 740, as compared to the number running on port 80(1088). Furthermore the first part of the hypothesis was also incorrect, the number of codes which makes a server inaccessible is less than the ones which made them accessible. The information breakdown can be seen below.

<table>
<thead>
<tr>
<th>Accessible</th>
<th>Non-Accessible</th>
</tr>
</thead>
<tbody>
<tr>
<td>200 OK</td>
<td>434</td>
</tr>
<tr>
<td>302 Found</td>
<td>488</td>
</tr>
<tr>
<td>307 Temporary Redirect</td>
<td>8</td>
</tr>
<tr>
<td>302 Object Moved</td>
<td>47</td>
</tr>
<tr>
<td>301 Moved Permanently</td>
<td>148</td>
</tr>
<tr>
<td>302 Temporary Moved</td>
<td>6</td>
</tr>
<tr>
<td>302 Redirect</td>
<td>13</td>
</tr>
<tr>
<td>302 Redirecting</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>1148</td>
</tr>
<tr>
<td>401 Unauthorized</td>
<td>82</td>
</tr>
<tr>
<td>404 Not Found</td>
<td>192</td>
</tr>
<tr>
<td>403 Forbidden</td>
<td>253</td>
</tr>
<tr>
<td>500 Internal Server Error</td>
<td>51</td>
</tr>
<tr>
<td>400 Bad Request</td>
<td>13</td>
</tr>
<tr>
<td>400 Proxy Error</td>
<td>50</td>
</tr>
<tr>
<td>404 Error</td>
<td>17</td>
</tr>
<tr>
<td>503 Service</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>663</td>
</tr>
</tbody>
</table>

This research can help law making bodies understand that even with many of the healthcare devices which are visible to the public, only a minority of them are protected with a higher level of security and are inaccessible. However, it should be kept in mind that these devices may just be a very small subset of the actual amount of devices which are out there. Furthermore, these devices could be serving the purpose of honeypots, allowing people to intentionally attack vulnerabilities. However, based on the data which is present, the majority of the devices which Shodan found for healthcare related devices are open, accessible and are using the less secure port 80.
Section 6: References


