Hacker Influence and SCADA Devices – A Temporal Shift Approach to Risk Identification

MIS 510, Cyber Security Project, Spring 2014

Michael Byrd, Eric Case, Bradley Dorn, Wenli Zhang
Contents
Introduction/Background .................................................................................................................. 2
Cyber Security Research Questions ............................................................................................... 2
Research design ............................................................................................................................ 3
Data Collection ............................................................................................................................... 3
Analysis ........................................................................................................................................ 4
   Hacker Influence – Network Centrality ...................................................................................... 4
   Topic Analysis – Hacker Profiles ............................................................................................... 5
   Topic Analysis – Important Forum Topics ................................................................................. 6
   Connection of SCADA devices ................................................................................................. 6
Findings and Discussion .................................................................................................................. 6
   Hacker Influence ....................................................................................................................... 6
   Important Forum Topics .......................................................................................................... 8
   Connection of SCADA devices ................................................................................................. 9
Conclusions .................................................................................................................................. 10
References ..................................................................................................................................... 10
Introduction/Background

The primary focus of our project is shifts in influence of individual hackers. We use Hackerweb and Shodan to investigate antecedents, incidences, and outcomes of shifts in influence of hackers.

Cyber Security Research Questions

Network centrality has been used to identify critical leaders in hacker social networks (Lu, 2009). Finding these leaders can help law enforcement agents address issues caused by the networks to which these leaders belong. Beyond identification, more meaningful responses can be made if a deeper understanding of the nature of each leader’s influence can be established. One uninvestigated quality of this influence is the timing of its creation. With our first research question, we aim to use shifts in network centrality as indicators of increases of influence.

**Research question 1:** Can shifts in network centrality be used to identify the timing of increases of influence for hackers?

Timing of increases of influence for hackers adds context to the influence of each hacker. By tying time periods to influence increase, we can associate historic context information with the increase of influence. One specific piece of historic information that can procedurally be associated with these influence creation points is the set of topics that was being discussed by the community and the individual hackers at the time. With sufficient data, it may be possible to identify the topics that made the hacker influential. With the notion that these topics become a part of the profile of features for which the influential hacker is known, we investigate the automated creation of hacker profiles using these topic analyses in our second research question.

**Research question 2:** Can topic analysis during periods of increasing hacker importance be used to develop representative hacker profiles?

The topics that hackers discuss while becoming influential are interesting not only because they describe something about the hacker, but also because the subject matter may have significant societal implications. Our third research question takes an exploratory approach to investigating the topics discussed by hackers while becoming influential.

**Research question 3:** Can individual shifts in influence be better understood and can trends of broader significance be found by investigating topics frequently discussed by influential hackers?

In addition to shifts in hacker influence over time, we also sought to examine how the temporal dimension impacts other threats with significant potential for societal impact. One example of a security threat with significant societal impact was the Stuxnet worm. The worm was discovered in mid-2010. Its discovery sparked a flurry of security research into and awareness of Supervisory Control and Data Acquisition (SCADA) networks and devices (Gross, 2011). SCADA networks and devices are used to control and monitor industrial processes and also used to control and monitor things closer to home like traffic lights and street lights like those in Westminster (Seifert, 2012). It has been three and a half years since Stuxnet and we wondered if SCADA operators had learned to protect their devices.

A search of IT security practitioner literature found many examples informing SCADA operators and IT security practitioners how to find and protect SCADA devices. A few of the better examples are the Open Web Application Security Project (OWASP) has a SCADA Security Project.
(http://www.owasp.com/index.php/OWASP_Scada_Security_Project), the ISGroup’s SCADA Exposure website (http://www.scadaexposure.com), both of which raises awareness of exposed SCADA devices, and of course, the Industrial Control Systems Cyber Emergency Response Team (ICS-CERT) from the Department of Homeland Security. ICS-CERT provides a great deal of information on protecting SCADA devices, but does not offer guidance on gaining visibility into the network.

Accordingly, our fourth research question addresses these concerns using Shodan to monitor the frequency of SCADA devices connections to the Internet. A successful attack on these devices could affect a whole city at once unlike a successful attack on a large merchant, which could affect the same number of people but spread across the country or globe.

**Research question 4:** Can trends in the connection of SCADA devices to the Internet be identified and monitored through the examination of port crawling records as captured by Shodan?

**Research design**

**Data Collection**

We gathered data from a subset of the forums contained in the Hackerweb database for analysis. We believe the forums we have selected are generalizable to the entire set of forums. Our goal was to select forums with each of three primary languages: English, Russian, and Chinese with comparable data. We first generated the count of posts each day for each of the 20 forums for the lifetime of each forum. After inspecting these counts, we chose to use the data for the 2010 year because it had sufficient data across each month for at least one forum of each language. We then selected three forums that appeared to have representative post counts from the complete set. The forums we selected are VCTool, Unpack, and Antichat. Figure 1 represents the results of counting the posts by day for each forum,

![Posts Over Time Graph](image_url)

**Figure 1: Hacker forum post frequency**
Data about hackers and the content they contribute can be easily queried from the Hackerweb database. We utilized SQL scripts to extract post and user information for the selected forums for the selected date ranges. Queries were created manually in the Shodan tool to identify connected SCADA devices. The results of these queries were exported via a python script to a spreadsheet for analysis. These queries and scripts could be repeated automatically in the future.

The data we collected was cleansed and normalized prior to analysis. Date formats in the Hackerweb data varied widely. We generated custom SQL and Perl scripts to normalize the date formats for analysis (Friedl, 2006). For the topic analyses, it is important that equivalent concepts are treated as equivalent even when their form may slightly differ. Some normalizing steps we took to ensure this are: conversion of all text to lower case, removal of URLs, standardize the reference of other users (for example, @username linking (Fox, 1989)), removal of hashtags, removal of unnecessary punctuation, removal of excess punctuation, removal of uninformative “stop” words (such as a, is, the, with, etc.) (Fox, 1989; Zou, 2006) automatic fixing of some typos (for example, changing “hunnnngry” to “hungry”), and removal of words that don’t start with letters.

Analysis
After collecting and cleaning our data, we calculated centrality measures for every hacker on our subset of forums for every month of 2010. We then performed our topic analysis and analyzed at the individual hacker level and at a forum level. Finally, we analyzed the Shodan data and sought to identify commonalities within the two temporal data sets.

Hacker Influence – Network Centrality
To measure the influence of a hacker over the community of hackers to which they belong, we use network centrality. Shifts in influence are identified by shifts in network centrality. A relationship between two hackers is considered to exist if they have both contributed to the same thread.

Network centrality is a measure of “the relative importance of a node within a graph” (Newman, 2010). There are various measures to determine this ranking, such as degree centrality” (Lu, 2009). Another measure of centrality is Eigenvector centrality. Eigenvector centrality is a measure of the influence of a node in a network (Ruhnau, 2000). It assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes” (“Centrality - Wikipedia, the free encyclopedia,” 2013). Google’s PageRank is a variant of the Eigenvector centrality measure.

These calculations require some mechanism for associating nodes. For our analysis, we considered nodes connected if they have contributed to one or more threads together. After making associations between nodes, we calculated the eigenvector centrality for each hacker as a proxy for their influence within the network.

We then selected five users with large shifts in centrality for each of the three forums we assessed. Figures 2, 3, and 4 show the results of our influence assessment across the three forums we analyzed for the year 2010 for the selected user IDs. Grey squares represent periods of low influence preceding periods of high influence. Periods of high influence (proxied by centrality) are indicated by blue squares.
Without additional context, it is difficult to determine whether our results have successfully identified the most influential hackers. We do know that we have been able to identify hackers with the highest number of associations with other hackers, but that is limited as a metric of influence because we don’t know the nature or successfulness of the interactions. To gain this additional context, we next examine the content of the posts using computational linguistics techniques.

**Topic Analysis – Hacker Profiles**

After identifying periods of significant changes in influence, we use computational linguistics to perform a topic analysis of discussions in which each newly influential hacker participated. We identify topics of discussion before, and after the shift in influence separately. These topics are then used to form a hacker topic profile. This profile is based on the assumption that each hacker will become associated with the topics that make them influential, similar to how academic writers become associated with their influential work.

Our topic analysis looked for words and word-pairs that occurred the most frequently across the postings in the relevant set (i.e. user/month) (CoreNLP, n.d.; Huang, Qi, & QIAN, 2007). We also
considered the association rule. With the occurrence of certain words, it generally follows that another associated word will also be in the document (Agrawal, Imieliński, & Swami, 1993; Wang, Tang, Han, & Liu, 2002). For example when premises are in a document, the word conclusion generally follows. We looked for the most frequently associated set of two-word combinations.

For each hacker with a significant shift in influence, we performed a topic analysis for their posts the month prior to their increase in influence and the month after their increase in influence. We then manually compared the most highly ranked words and word-pairs in an attempt to discover what it was that made them influential. The results of these comparisons are discussed below.

**Topic Analysis – Important Forum Topics**

In reviewing the topics extracted for particular users, it became apparent that there were common topics across the users. In order to evaluate the users better we decided that a baseline set of topics would be useful. To generate this baseline, we performed a similar assessment of topics on the forums as a whole. After cleaning and normalizing our data, we similarly extracted the top topics by month for the five or ten users with the overall highest centrality in each of the three forums. We then reviewed these topics in conjunction with the individual-shift results. Results are discussed below.

**Connection of SCADA devices**

In 2004, Johnny Long’s book “Google Hacking for Penetration Testers” showed how hackers could use web search engines to find valuable information exposed on the Internet as well as interesting targets. It also showed security practitioners how to gain insight into their networks and the behavior of their users. In 2009 John Matherly created Shodan, a database of Internet connected devices beyond web servers. Shodan takes Long’s idea further by looking for ports other than port 80. With Shodan, security practitioners can see printers, routers, cameras and other devices on their networks such as industrial control system devices like those targeted by Stuxnet.

Using search terms from the OWASP SCADA detection cheat sheet, work done by Francesco Ongaro and Gianluca Pericoli of the ISGroup SRL (2013) and terms offered by a pastebin user (Guest, 2012), we were able to find many SCADA devices in Shodan. Given that Shodan has collected information about the devices we are interested in studying, the question became how to analyze the data in Shodan. Shodan’s data can be accessed via its application-programming interface (API), which can be accessed from within Google’s Spreadsheet application (Matherly, 2014b).

During our analysis, we noticed 21 months that always returned zero results. We learned that Shodan was moved to a new architecture in January 2013 and historic data was only partially imported (J. Matherly, personal communication, February 22, 2014). We decided to focus our analysis on the twelve months, February 2013 through January 2014, which meant simply changing our spreadsheet from quarters to months.

**Findings and Discussion**

**Hacker Influence**

Performing a topic analysis before and after the top identified hackers gained influence revealed very distinct topics when comparing the two periods. Figures 5, 6, and 7 represent our findings for each of the three analyzed forums. A key weakness that we have identified to our approach is that user may
have low centrality scores due to a small number of posts. In future research, a mechanism to compensate for this limitation would need to be developed. For those users that did exhibit shifts from one set of topics to another, we identified several trends. Two users appeared to shift from technical questions about how to access software and tools to how to use those tools and sharing them with others. Other users shifted from a content generation and consumption role to more administrative topics, possibly resulting in additional social connections, leading to higher centrality. However, another user in the Chinese forum shifted to a more content focused role – this is a finding the most congruent with our research questions, this user’s content area could potentially have driven his influence. The content appears to be related to tool acquisition and shell scripting.

<table>
<thead>
<tr>
<th>UserID</th>
<th>Pre-shift Month</th>
<th>Number of Posts</th>
<th>Pre-shift Topics</th>
<th>Post-shift Month</th>
<th>Number of Posts</th>
<th>Post-shift Topics</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>UID 2289</td>
<td>March</td>
<td>22</td>
<td>server, use, ircd, tutorial, connect</td>
<td>April</td>
<td>75</td>
<td>freeboot, file, megaupload, xbreoot, rar, part, rapidshare, bin</td>
<td>Appears that this user got involved in file sharing heavily in the month where his centrality went up</td>
</tr>
<tr>
<td>UID 2688</td>
<td>March</td>
<td>-</td>
<td>/</td>
<td>April</td>
<td>52</td>
<td>windows, bit, hub, time, wrote, connection, failed</td>
<td>Centrality increase is likely an artifact of the limited number of posts in the preceding month</td>
</tr>
<tr>
<td>UID 3461</td>
<td>May</td>
<td>48</td>
<td>frozen, admins, elitbot, c0ntr0lpunk, poor, information, use, name, credit</td>
<td>June</td>
<td>139</td>
<td>file, server, download, site, bot, address, channel, code, eax, web, password, bandwidth</td>
<td>This user appears to have shifted from a focus on hacking tools to distribution of tools and software</td>
</tr>
<tr>
<td>UID 516</td>
<td>January</td>
<td>2</td>
<td>work</td>
<td>February</td>
<td>20</td>
<td>link, greylink, provide, crash, cod, dat, btw, client, provide_wid</td>
<td>Centrality increase is likely an artifact of the limited number of posts in the preceding month</td>
</tr>
<tr>
<td>UID 6466</td>
<td>July</td>
<td>2</td>
<td>doesn't, doesnt_work</td>
<td>August</td>
<td>55</td>
<td>aimbot, download, cracked, thanks, virus_c_c, help, version, cracked_version</td>
<td>Centrality increase is likely an artifact of the limited number of posts in the preceding month</td>
</tr>
</tbody>
</table>

Figure 5: Topic Analysis for Influential Hackers on VCTool (English)

<table>
<thead>
<tr>
<th>UserID</th>
<th>Pre-shift Month</th>
<th>Number of Posts</th>
<th>Pre-shift Topics</th>
<th>Post-shift Month</th>
<th>Number of Posts</th>
<th>Post-shift Topics</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>UID 10809</td>
<td>June</td>
<td>42</td>
<td>注册</td>
<td>July</td>
<td>112</td>
<td>注册, 模式, 功能</td>
<td>User appears to provide assistance to new users. It seems that influence increased as task complexity increased.</td>
</tr>
<tr>
<td>UID 1420</td>
<td>June</td>
<td>42</td>
<td>注册后, 支持, 感谢, 登陆</td>
<td>July</td>
<td>99</td>
<td>注册后, 原帖, 原帖, 发表</td>
<td>User appears to provide descriptive information about posts. Influence increase likely associated with post-count increase.</td>
</tr>
<tr>
<td>UID 13881</td>
<td>June</td>
<td>74</td>
<td>原帖, 发表, 去壳</td>
<td>July</td>
<td>103</td>
<td>朋友, 注册, 大哥, 注册后, 模式</td>
<td>User appears to shift from a content generating user to a more social or moderating role, resulting in more connections.</td>
</tr>
<tr>
<td>UID 3</td>
<td>June</td>
<td>133</td>
<td>下载, 享更多, 信息保护, 软件加壳, 加密, 多功能, 安全, 密码, 游戏, 病毒, 编译</td>
<td>July</td>
<td>212</td>
<td>下载, 保护, 注册后, 进程, 安全, 教程, 插件, 程序, 软件安全</td>
<td>User appears to shift from a content generating user to a more social or moderating role, resulting in more connections.</td>
</tr>
<tr>
<td>UID 31361</td>
<td>August</td>
<td>47</td>
<td>下载, 功能, 支持, 注册, 登陆</td>
<td>September</td>
<td>239</td>
<td>下载, 注册后, 功能, 登陆, 加壳, 内壳</td>
<td>User appears to shift from a community support role to a content generation role while continuing community support.</td>
</tr>
</tbody>
</table>

Figure 6: Topic Analysis for Influential Hackers on Unpack (Chinese)
Important Forum Topics
To provide context to the individual shifts in topics and to look for broader trends, we also extracted topics from the top five or ten users by centrality for six months. We used five and ten for the Chinese and Russian data sets due to the substantially greater number of posts such that our computations would complete in a reasonable amount of time. The results are presented in Figure 8.

We make the following observations from this result set: First, there is an obvious permeation of administrative topics. Many users are discussing logistic details like how to post and register. Future research should work to filter these types of posts out. Second, there are a large number of topics...
related to the acquisition of software. While this could potentially be interesting as an area to explore independently, it would probably make the results of our individual-shift analysis more interesting if these types of topics could be filtered out. This would lead to more emphasis on the actual type of hack where methods and targets might be discussed. For example, instead of logistical issues of acquiring tools. Finally, we note evidence of an external event impacting the discussions on the forum: the 2010 earthquake in the Sichuan province in China. This illustrates the connection between forums frequented by hackers with the broader social context. Information on the earthquake was heavily censored in the mainstream Chinese media, the hacker forums are more difficult for the state to control or censor and as such may provide an important parallel mechanism for social discourse.

**Connection of SCADA devices**

Twelve months of data was enough to show SCADA practitioners are continuing to attach their devices to the Internet – creating the “Internet of Things.” The first device we looked for was the type of device targeted by Stuxnet, a Siemens programmable logic controller (PLC). Shodan discovered over 700 of them this month (February 2014), but does not have historic data for them. At the other end of the spectrum is VxWorks, a real-time operating system – over 62,000 discovered this month.

We removed both of these from our analysis and selected four collections of devices:

1. Moxa products are used “in a wide range of applications, including factory automation, smart rail, smart grid, intelligent transportation, oil & gas, marine, and mining” (“Moxa Company Profile,” n.d.).
3. IPC@CHIP products from Beck IPC GmbH are “for all aspects of industrial control” (“Company profile,” n.d.).

Each collection of devices showed an increase in devices Shodan discovered each month. Projecting a linear growth rate, their growth rates ranged from 119/month for Moxa to 290/month for WindWeb (see figure 9).
These data shows three and a half years after Stuxnet, practitioners continue attaching SCADA devices to the Internet. However, it does not show if the devices are secure or not. We should not assume they are secure, as the recent news about a backdoor in several wireless routers (Matherly, 2014a) shows devices once thought secure can have new flaws discovered.

**Conclusions**

Identifying hackers and providing context to their most influential actions is a meaningful and automatable exercise. In this study we have demonstrated the possibility of conducting such an analysis in an automated fashion. While our study is subject to limitations given the timeframe and data set, the methodology could easily be extended in a more robust exercise. We believe our set of results is indicative of the types of results that would be found with other similar data sets. The techniques used in this study can also be applied with little modification to data collected about hacker communities of any type to enhance the generalizability of the findings.

In summary, we found that shifts in network centrality can be used to identify the timing of increases of influence for hackers (Research question 1). Topic analysis during periods of increasing hacker importance can be used to develop representative hacker profiles (Research question 2). Also, that individual shifts in influence can be better understood and can trends of broader significance can be found by investigating topics frequently discussed by influential hackers (Research question 3). By examining data from Shodan, we were able to identify and analyze temporal trends in the connection of SCADA devices to the Internet (Research question 4). Taken together these results illustrate the power of examining temporal trends in data – not simply to identify changes in the data itself, but as a tool to better understand the dynamic nature of human and information systems.
References


Wang, K., Tang, L., Han, J., & Liu, J. (2002). Top down fp-growth for association rule mining. Springer.