**Introduction**

The internet has always been distinguished for its ability to foster interactions and discussions. Criminals have taken advantage of these features to maximize their criminal opportunities. Newcomers can now get advice from experienced criminals for their technical expertise. Cyber analytics provides an opportunity to understand the dynamics of the hacker community and the vulnerabilities of information systems. In this project, two security testbeds, Shodan and Hacker Web, will be utilized together to identify a threat and the available threat landscape. Shodan is a search engine that allows you to search for specific systems and internet enabled devices. The hacker web contains forum postings from international hacker communities, developed by the AI Lab at the University of Arizona.

Multiple aspects of cyber security can be revealed through the interaction of the hacker and shodan communities. The first step was to find the more followed members of the hacker community, based chiefly on their reputation scores composed of other factors in the forums. Then we see what topics, sites and tools most members are interested in. Through Shodan, we can search for an inventory of systems that could be targeted by these tools, hackers or blacklisted sites. Finally we analyzed whether the use Shodan could produce a vulnerability report based on our previous findings.

**Literature Review**

1. **Shodan**

   Shodan’s ability to uniquely identify systems is shown in the 2011 paper by Eireann P. Leverett. Mr. Leveret was able to uniquely identify over 30 systems using Shodan, including their location, manufacturer, model, type, and purpose. It took over two years for to collect and analyse and was able to produce findings for 7489 units [6]. Over the last several years the the functionality and scope of Shodan has increased. This has decreased the technical knowledge required of users to uniquely identify a system.

   While Shodan has decreased the entry level for many users, specific domain knowledge is still required to get meaningful results. A beginning user would type SCADA and they would find 372 listings [9]. However, an experienced user would search for more specific terms like Modbus, Simatic+HMI, and S7 [7]. The experienced user can find specific systems with certain attributes.

2. **Hacker tools and target systems**

   Most Black hat hackers build custom tools to launch anonymous attacks. However, most beginners, namely script kiddies, start with existing tools. Hacker Web provides us with an insider’s glimpse into this underground hacker community.

   Andy O’Donnell, a well-reputed security enthusiast listed Nmap, Metasploit, Wireshark as some of the the high-profile tools in BackTrack Linux’s arsenal [10]. Many other security experts such as Gordon ‘Fyodor’ Lyon have included these tools amongst the top/most used hacker tools [11][12]. Mr. Lyon, the coder of Nmap has pioneered the tool to exploit and break into net devices. He has given extensive demos at DefCon on the subject, not even Microsoft’s own intranet was safe from his “browsings”. His latest project surveyed 3,000 people about the latest top security tools and relaunched the Insecure.org site as SecTools.Org [8]. This new site socially and dynamically keeps track of reviews and ratings of the latest security software and tactics.
Armed with cross-verified lists of security tools, we asked what could be some attractive soft targets? Shodan provides us with access to a range of systems and their specifications, for instance their ports and server operating system they run on. Windows and Linux have a major market share for server operating systems. Apache’s proliferation as a widely used open-source, reliable platform for ⅔ of the internet’s surface websites makes it crucial in our analysis. Exploration of data related to the hacker tools and their targets in Hacker Web lead to some interesting discoveries.

III. Hacker characteristics & interests
Skilled hackers fall into three categories or ‘hats’ based on their motives. White hats, or ethical crackers, use their skills to penetration-test and reveal security holes, briefing institutions on how to harden their systems [14].

Black hat hackers are usually associated with malicious intent and deliberately cross the legal line. Their main interests used to be predominantly ego-gratification, have shifted to highly lucrative financial gain and the theft or hijacking of sensitive personal and political information.
Grey hats are a mix of the two and walk on a thin legal line, trying not to cross it. An example is a researcher who gained access to the “keys of the Facebook kingdom” through a remote execution flaw. He reported the bug to Facebook who paid him $33,500 for the disclosure.
There also exists hobby hackers and game hackers whose main motives are to increase the ability of the device, to give it a feature or just to learn [14]. A classic example is cofounder of Apple, Steve ‘Woz’ Wozniak. Before it became the most successful company on the planet, it started with him and Steve Jobs tinkering in their garage and showing their creations to the local computer homebrew club. A modern-day successor who went from a Ph.D at Carnegie Mellon, to showing a head-tracking 3d system using Nintendo Wii on YouTube, to TED, to Microsoft’s Kinect and finally Google’s ATAP team about to release Project Tango. We’re talking about Johnny Chung Lee.

ROM hackers edit ROM files directly. This kind of crime usually involves social engineering and utilize tactics such as vanity, honor, helpfulness, authority, etc. One of the most common interests of hackers includes dictionary attacks with rainbow tables, brute force attacks and keylogging targeting weak passwords. Most common techniques for hacking include injection, cross-site execution, vulnerability scanning, password-cracking, packet-sniffing, spoofing, phishing, rootkits, social engineering, trojans, remote access, computer viruses, worms and keyloggers.

IV. Significance of Hacker reputation
Evidence shows that less skillful criminals obtain help from experienced hackers and organized crime groups since this makes them less likely to get caught [2]. Such hackers have a reputation within their communities as they are relied upon by their fellow hackers. Researchers have acknowledged that the reputation for leadership is an important asset in maintaining relationships in social networks and among peers, superiors and subordinates. From a social network perspective, leaders are social architects who build and manage their ties with others to enhance both the performance of their communities and their personal reputation [3]. They are considered experts in their group and are expected to be knowledgeable in the field.

Trust is a key factor among members of cybersecurity forums. The reputation score of these authors play a big role in deciding whether an author is trustworthy or not. Leadership status is closely tied to high rates of contribution - levels of activity in the community. This helps them increase their
reputation and trustworthiness [4]. The hierarchical chain within cyber criminal groups is linked to an author's reputation and hence is a valuable attribute hackers desire and pursue [1]. Prior related works from Benjamin et al have explored two major hacker communities from the US and China with the aim of determining key actors. Two features, average message length and number of replies per thread, were studied to understand the hackers discussion quality. Four other features, number of threads involved, tenure, sum of attachment and total messages were used to model the extent of involvement in the group [1]. A review of these past studies gives us an idea of various factors that could possibly contribute to the overall reputation scores.

**Research Questions**

**Shodan**
1. Can shodan uniquely identify system types?
2. Can shodan be used to assess vulnerability to hacker software?
3. Where are the top blacklisted hosts located around the world?

**HackerWeb**
1. Are hackers on the forums using the most popular tools and operating systems?
2. What are the top websites and attachments shared on the forums?
3. When are hackers posting the most throughout the year down to the hour?
4. What are the forum members top interests by level of expertise?
5. What calculated attributes were used to derive Author’s reputationScore?

**Research Design**

**Tools**
- DB: HeidiSQL, MySQL
- IDE: Eclipse, PyDev, Java & Python connectors & language
- Analysis & Visualization: Microsoft Excel PowerPivot, Prism 5, RapidMiner, Tableau

MySQL programs were connected remotely to HackerWeb via VPN, java/python connectors and IP addresses with authentication. Data was transferred to analytics and visualization programs via CSV files. Shodan’s API was used to connect to it using the class key mapped to Python IDEs PyDev and Eclipse. Python script was written employing SQL queries to search the online database.

**Assumptions**
Given the time, knowledge, and computing constraints, we focused our analysis of the forums based on the following factors:
- Impact: U.S. is home to more malicious attacks than the rest of the world
- Familiarity: with the English language, culture, and domain
- Spread: the variability of the data sufficient when uniting tables
- Practicality: how difficult it is to clean, aggregate and process data into meaningful information
- Precision: could we narrow down the measurements to find significant patterns (ie. regression)
- Speed: the foreign language forums were ten to twenty times bigger in size (20-40 MB vs. 400-900 MB)
The non-programming languages spoken in the forums were connected with their respective geographic regions like China, Russia, Persian or Middle East. We understand that all the forums could be queried faster using more powerful cloud computing such as Amazon EC2 and Hadoop.

**Issues**

**Unknowns:**
- bleeding-edge hacking methods or exploits
- how the dark web community operates
- what custom code and tools are used
- what fields they operate in or types of targets they go for
- hacker expertise level and forum composition of them

With the database itself:
- all data saved in “varchar” format as opposed to numbers being saved as dates or integers.
- field formats were different across the tables (ie. 24 January 2013 vs. 01-24-2013)
- some fields are null or contain different ranges (ie. dates from non-overlapping years)
- content is inconsistent (non-standardized layout, unstructured text, links, etc.)
- do not have permission to create/modify/run queries on database such as `match against()`
- not completely normalized, but good enough

For example, 22 would show up as being less than 3 because it was being ordered by the leftmost digit as if alphabetical. Comma-delimited numbers in the thousands were seen by MySQL as decimal points.

**Methods**

We wrote queries to browse content, get avenues of approach as to how feasible questions were to answer and how to precipitate patterns. Tables were tied together, fields prioritized and weighed. Transformation of varchar strings to appropriate number formats and consistency checks were carried out. Dates were converted to date-time formats and segregated into individual units - year, month, day, etc using `STR_TO_DATE` function. To find *Tenure*, `DateDiff` was used to calculate the length of time since last posts and `Now()`.

Similarly, other numerical measures such as *postRank* or *numOfViews* were stripped of their commas (if in the thousands), cast as integers, sorted by number of digits and numerical values in descending order to achieve “natural sorting” of numbers. Nulls were changed to 0s using `Coalesce` and `NullIf`. For valid data, only positive numbers greater than 0 were allowed. After numbers were sanitized, calculations and extractions could proceed.

To extract links and attachment names from the content of forums, MySQL’s RegExp was insufficient. Instead, we resorted to finding keywords using `WHERE LIKE %`. From those rows we found reference points using the HTML tags in `contentWithHTMLTag` columns such as “href”. Once anchored by the reference, we could narrow down to the names using `Substring_Index`, trimming down to delimiting characters like “/” or “” in nested queries cropping from both side until arriving at the desired string. Since there were many kinds and formats of links, it was difficult to catch all of them. The ones that started outside of the alphabet were filtered out.
To find combinations of tools with operating systems mentioned, tables were aliased. Both outer and inner joins were used in sync to aggregate tables. To analyze the full set of forums, the UNION operator connected all the tables together. Once the analyses were carried out on the English tables, they were proof of concept of automation, applicability and scalability to other language forums. This is because many English words such as brands/trademarks or programming languages remain unchanged across the multilingual forums.

Analytics
To make inferences and make sure queries were addressing the questions with stated assumptions, random sample content was read from the flatContent of threads.

ReputationScore, postRank, numOfViews, numOfPost, numOfReplies measures including length of time on forums were used to separate the novice Authors from the more seasoned. Depending on the order and weights those attributes were sorted by, we deduced and tailored to different demographics and interests of forum members according to names of top threads. Greater influence was given to ranks by multiplying them so their numbers were comparable to the higher measures such as total numOfViews. In order to reduce statistical outliers, authors with a reputation score of 0 were not used in the analysis since these authors might have been inactive or new to the community. Bad reputation in mainstream population could also be caused by their origin, style and socioeconomic status [5] in the community.

To find website links, the “href” keyword caught most posts with links. They were consolidated by name, sorted by postRank (to find the top posts) and alphabetically, then counted the number of duplicates.

To find names of file attachments, the hasAttachment column proved to be too vague as it referred to things like picture links as well. So keywords like “attachment” and “views” were included. Since most posts with file attachments had standard content with “views” and file “size”, they were invaluable for extraction.

Many attributes were joined between tables, counted and sorted from high to low (descending). To discover different tastes in views or downloads, attributes were sorted in different orders with different weights and calculations.

To reverse engineer how a ReputationScore was derived, attributes were calculated in different permutations with averages, standard deviations and regression analysis using Prism 5 and Tableau. Outliers such as attachments and code length were quickly identified. Other fields closely related to posts were compared side-by-side showing similar regression lines to ReputationScore.

Correlation computed the value of the Pearson correlation coefficient r whose values ranges from -1 to 1. Correlation was chosen over regression since it makes no prior assumption as to whether a variable is dependent on the other. It gives us the degree of association.

The top tools were selected after going through various web resources and obtaining a list of the top 20 tools. To find the number of posts in which the tools were mentioned, a keyword search query was executed on the content of the posts. The tools with the maximum number of occurrences in the posts were selected for analysis. To find the degree of association of these tools with the target systems, a combined keyword search was performed for a count. For the sake of clarity, the axes of graphs were
tweak with respect to ranges and logarithmic scales while attempting to maintain representative integrity.

**RESULTS AND DISCUSSION**

1. **Can shodan uniquely identify system types?**
   Shodan can uniquely identify OS types as seen in Figure 1. It was able to detect about 63% of Linux Apache followed by Windows at 15%. Other features such as tool counts, maps and operating systems with details such as city, asn, organization, IP addresses and port numbers was also identified (refer to Figure 2).

2. **Are hackers on the forums using the most popular tools and operating systems? Can shodan be used to assess vulnerability to hacker software?**
   After running a keyword search through the forums, top tools like NMAP, Wireshark, NetCat and Metasploit were found to be most mentioned in the forums as seen in Figure 4. Also after a review of the potential target systems, the most popular ones were selected to be analysed along with the above tools. A count of the instances of co-occurrence of each of the tools and the targets was retrieved from the Hackerweb forums. The result had it that Windows and Metasploit were mentioned together in posts the most number of times as shown in Figure 3. Does that mean that Metasploit is most used on Windows to target other systems or should we consider that Windows is most susceptible to attacks launched through Metasploit? Could we be on the verge of uncovering all the vulnerable Windows systems/servers? A search in Shodan shows that there are close to 8.5 million open windows systems to launch an exploit.

3. **Where are the top blacklisted hosts located around the world?**
   The top 10 list was collated from a trusted DNS website. The IP addresses and domains were searched on Shodan, referencing the latitude and longitude in return. These were plotted on Tableau. The findings align with statistics of attacks from U.S. and China, but strange to find Austria on the map, South Africa, but not Russia, Persia, Middle East or some internet-liberal European states. Refer to the blacklist host map in Figure 5.

4. **What are the top websites and attachments shared on the forums?**
   Elitehackforums.com was listed 10 times more than the second place link. A lot of the website links were obscured by using an online link shortener such as adf.ly, bc.vc, goo.gl, and ge.tt; or a filelocker like mediafire, ge.tt, pastebin, multiupload, hotfile and rapidshare. Coming in third through fifth place were YouTube.com, MediaFire.com, and msn.MSdn.Microsoft.com. Pictures have the highest views by a staggering margin, primarily in the “General Discussions” subforum as shown in Figure 8. A file called main.png usually known as a picture was top of the list. The forum member describes it was being nice code, but incomplete by someone with the Twitter handle of MescalineRC. The last Tweet refers to a comment about streaming webcam video. This seems to suggest it’s some code or picture having to do with remote access to webcams.

   The top subforum with the highest number of downloads for individual files is “Trojans; Remote Malware”; and the one with the greatest total volume was “Delphi”. Borland Delphi is a tool for developing Windows 32 bit programs with a gui or DOS. The second file is WaterEffect.rar. It is source code for creating banner that simulates water use in forum profiles. BlackNix is a ‘rat’ coded...
in Delphi with many powerful features as a remote-connect bot that can bypass defenses of a computer. It can act as a keylogger, report computer stats, run shellcode, be a file services, processes, registry, application, and Wireless Manager.

5. **When are hackers posting the most throughout the year down to the hour?**
The time with the highest activity after aggregating tables grew steady from 2008, to 2013 with peaks and troughs according to standard rates of activity throughout the seasons and the week (i.e. less before New Years and mid-week). The busiest month was August on Saturday nights at 1 a.m (Figure 10 & 11).

6. **What are the forum members top interests by level of expertise?**
Across most rankings, gaming guides like “Free Modern Warfare 2 Hacks+More!” was always highly rated as was Minecraft. Hacker tools were next “Ultra Hacker Pack” and “Hacker Underground Handbook”, as well as proxy tools. When ordered by views followed by rank, the second post was an “MSN Mirc Bot”-has to do with freezing, DDOS, mass adding, etc. The most viewed post but not as highly ranked was how to setup a Zeus Bot (the derivative called SpyEye was used to steal millions of credit card numbers from Target) as seen in Table 1.

Members like music as seen with top author Mi4night’s post “Good Music” the prominence of YouTube and mp3 site playlists. There was a marked interest in hacking into accounts like Facebook, Skype or Gmail; also to get free memberships to porn sites and hacker forums.

7. **What calculated attributes were used to derive an Author’s reputationScore?**
Although the results of the Pearsons correlation portray that the relationship between various factors and reputation score is significant, only a Pearsons r of 0.7 - 1 is considered strong. The results suggests that features such as number of posts, total number of replies, total number of views, message length influence the reputation score more heavily than the other factors. Thus, we can infer that the quality of the forum i.e. the amount of contributed content may have an effect on the reputation score. Code length and attachment were missing on most posts and hence a valid analysis could not be performed. From Table 2, we can also see that the impact of the different factors vary between forums. For instance, these factors had a big impact on forums such as Hackhound and unpack compared to others. Regression curves from Figure 12 shows the trend of the aggregation of all considered factors against reputation score.

Studies have suggested that reputation evolves through time and is affected by behavior [5]. Surprisingly, tenure did not have a significant correlation with reputation score on most of the forums. Research shows that the reputation score can be categorized into three classes: characteristics, networking and behavior[5]. Other factors such as criminal mentors, number of contacts, the structure of personal networks, experience and skills low self-control and enrollment in criminal organizations may also have an impact [5]. This paper is again a reminder that there are other fields that can be studied to understand social behavior such as recognition and market dynamics. In summary, the level of forum activity and consistency are important to convince others of their trustworthiness and build reputation.
REFERENCES

APPENDIX

SCREENSHOTS SHODAN

UNSECURED SEVER PORT 80

230 10.1.1.11 FTP server ready
230 Anonymous access granted, restrictions apply.
214 The following commands are recognized (* = not implemented):
CDW SCDP SCUP SHUT QUIT PORT PASV
PASS EPSV ALLO RNFR RNTO DELE MDMD
XMKD XMKD PWD XMKD SIZE SYST HELP
NOOP FEAT OPTS AUTH CCS COMP ENC MDC
PBS2 PROT TYPE STRU MODE RETR STOR STOU
APPE REST R...

UNSECURED FTP

213.233.183.49
StarFi University Of Technology
Added on 25-09-2013

User Access Verification

Password:

PRAGUE WEBCAM
Figure 1. Server operating systems scanned by Shodan

Figure 2. Range of specifications retrieved from Shodan.
Figure 3. Most popular tools and operating systems used in various forums retrieved from Hacker Web.

Figure 4. Mention count of hacking tools used across various OS. Data obtained from Hacker Web.
Figure 5. Top blacklisted hosts located around the world.

Figure 6. Top website attachments shared on forums by frequency and size.
Figure 7. Top website attachments by popularity shared in sub-forums.

Figure 8. Sub-forums categories where attachments are generally shared.
Figure 9. Top websites mentioned on forums.

Figure 10. Activity density of hackers by time of day.
Figure 11. Activity density of hackers by month.
Table 1. Forum members top interest by level of expertise.

<table>
<thead>
<tr>
<th>Title: By Highest Views &amp; Post Rank</th>
<th># Views</th>
<th>post Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free Modern Warfare 2 Hacks + More!</td>
<td>12,509</td>
<td>63</td>
</tr>
<tr>
<td>[DOWNLOAD] Msn Mirc Bot (Freezer, DDos, Mass Add + More)&lt;/s&gt;</td>
<td>12,395</td>
<td>42</td>
</tr>
<tr>
<td>[Tutorial] Setup Zeus Bot WITH Pictures &amp;a; Tutorial</td>
<td>62,039</td>
<td>5</td>
</tr>
<tr>
<td>[Must READ] Proxies (Socks, Http, ect..) [KEEP ADDING]</td>
<td>27,509</td>
<td>9</td>
</tr>
<tr>
<td>Brian's Online eMail Bomber</td>
<td>19,412</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Title: By Top Posts &amp; Authors</th>
<th>post Rank</th>
<th>Author Name</th>
<th>Reputation Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good Music</td>
<td>99</td>
<td>Mi4night</td>
<td>5</td>
</tr>
<tr>
<td>Ultra Hacker Pack</td>
<td>99</td>
<td>pack</td>
<td>0</td>
</tr>
<tr>
<td>The Hackers Underground Handbook.pdf</td>
<td>99</td>
<td>blackeagle</td>
<td>0</td>
</tr>
<tr>
<td>The Big List Of Accounts</td>
<td>99</td>
<td>RiskyBusiness</td>
<td>0</td>
</tr>
<tr>
<td>How To Hack A Skype Account [Very Easy&amp;#33;]</td>
<td>99</td>
<td>Rickman</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Title: By Rank, Posts, Views - even distribution</th>
<th>post Rank</th>
<th># Posts</th>
<th># Views</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook working Hack! private method-InTrUd3r</td>
<td>137</td>
<td>234</td>
<td>6,937</td>
</tr>
<tr>
<td>Get your first award on EHF</td>
<td>69</td>
<td>126</td>
<td>5,865</td>
</tr>
<tr>
<td>Official Tool- Proxy Finder</td>
<td>66</td>
<td>104</td>
<td>3,614</td>
</tr>
<tr>
<td>Free Modern Warfare 2 Hacks + More!</td>
<td>63</td>
<td>147</td>
<td>12,509</td>
</tr>
<tr>
<td>FREE Elite membership Elite Hack Forums</td>
<td>46</td>
<td>53</td>
<td>2,721</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Title: By Rank, Posts, Views - Weighted</th>
<th>post Rank</th>
<th># Posts</th>
<th># Views</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook working Hack! private method-InTrUd3r</td>
<td>137</td>
<td>234</td>
<td>6,937</td>
</tr>
<tr>
<td>Signature for those who asked!</td>
<td>35</td>
<td>35</td>
<td>831</td>
</tr>
<tr>
<td>Porn Accounts Free</td>
<td>15</td>
<td>94</td>
<td>948</td>
</tr>
<tr>
<td>Get your first award on EHF</td>
<td>69</td>
<td>126</td>
<td>5,865</td>
</tr>
<tr>
<td>Free Modern Warfare 2 Hacks + More!</td>
<td>63</td>
<td>147</td>
<td>12,509</td>
</tr>
</tbody>
</table>
Table 2. Correlation Regression on reputation score using Prism 5. Assuming data are sampled from Gaussian populations (Pearson).

<table>
<thead>
<tr>
<th>Group</th>
<th>Goodness of fit</th>
<th>NumofPosts</th>
<th>Total#ofReplies</th>
<th>Total#ofviews</th>
<th>MsgLength</th>
<th>CodeLength</th>
<th>Attachment</th>
<th>Tenure</th>
<th>Aggregate of Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANON</td>
<td>Pearson r</td>
<td>0.526</td>
<td>0.364</td>
<td>0.600</td>
<td>0.726</td>
<td>0.378</td>
<td>-0.016</td>
<td>-0.045</td>
<td>0.731</td>
</tr>
<tr>
<td></td>
<td>Is the correlation significant?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R squared</td>
<td>0.277</td>
<td>0.132</td>
<td>0.360</td>
<td>0.526</td>
<td>0.143</td>
<td>0.000</td>
<td>0.002</td>
<td>0.534</td>
</tr>
<tr>
<td>ANTICHAT</td>
<td>Pearson r</td>
<td>0.414</td>
<td>0.364</td>
<td>0.157</td>
<td>0.277</td>
<td>0.119</td>
<td>0.243</td>
<td>0.111</td>
<td>0.286</td>
</tr>
<tr>
<td></td>
<td>Is the correlation significant?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R squared</td>
<td>0.172</td>
<td>0.133</td>
<td>0.025</td>
<td>0.077</td>
<td>0.014</td>
<td>0.059</td>
<td>0.012</td>
<td>0.082</td>
</tr>
<tr>
<td>ARHACK</td>
<td>Pearson r</td>
<td>0.573</td>
<td>0.307</td>
<td>0.213</td>
<td>0.444</td>
<td>0.254</td>
<td>0.488</td>
<td>0.020</td>
<td>0.237</td>
</tr>
<tr>
<td></td>
<td>Is the correlation significant?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>R squared</td>
<td>0.328</td>
<td>0.094</td>
<td>0.045</td>
<td>0.197</td>
<td>0.064</td>
<td>0.238</td>
<td>0.000</td>
<td>0.056</td>
</tr>
<tr>
<td>ASHIYANE</td>
<td>Pearson r</td>
<td>0.133</td>
<td>0.088</td>
<td>0.129</td>
<td>0.114</td>
<td>0.136</td>
<td>0.170</td>
<td>0.025</td>
<td>0.129</td>
</tr>
<tr>
<td></td>
<td>Is the correlation significant?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R squared</td>
<td>0.018</td>
<td>0.008</td>
<td>0.017</td>
<td>0.013</td>
<td>0.018</td>
<td>0.029</td>
<td>0.001</td>
<td>0.017</td>
</tr>
<tr>
<td>CINEMERARMAKY</td>
<td>Pearson r</td>
<td>0.295</td>
<td>0.025</td>
<td>0.084</td>
<td>0.303</td>
<td>0.346</td>
<td>0.388</td>
<td>0.090</td>
<td>0.099</td>
</tr>
<tr>
<td></td>
<td>Is the correlation significant?</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R squared</td>
<td>0.087</td>
<td>0.001</td>
<td>0.007</td>
<td>0.092</td>
<td>0.119</td>
<td>0.150</td>
<td>0.008</td>
<td>0.010</td>
</tr>
<tr>
<td>ELITEHACK</td>
<td>Pearson r</td>
<td>0.683</td>
<td>0.761</td>
<td>0.685</td>
<td>0.505</td>
<td>0.293</td>
<td>0.673</td>
<td>-0.235</td>
<td>0.586</td>
</tr>
<tr>
<td></td>
<td>Is the correlation significant?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R squared</td>
<td>0.467</td>
<td>0.579</td>
<td>0.469</td>
<td>0.255</td>
<td>0.086</td>
<td>0.452</td>
<td>0.055</td>
<td>0.343</td>
</tr>
<tr>
<td>EXPLICIT</td>
<td>Pearson r</td>
<td>0.636</td>
<td>0.487</td>
<td>0.485</td>
<td>0.447</td>
<td>0.283</td>
<td>0.351</td>
<td>0.043</td>
<td>0.475</td>
</tr>
<tr>
<td></td>
<td>Is the correlation significant?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R squared</td>
<td>0.404</td>
<td>0.237</td>
<td>0.236</td>
<td>0.200</td>
<td>0.080</td>
<td>0.124</td>
<td>0.002</td>
<td>0.226</td>
</tr>
<tr>
<td>HACKHOUND</td>
<td>Pearson r</td>
<td>0.795</td>
<td>0.790</td>
<td>0.809</td>
<td>0.783</td>
<td>0.334</td>
<td>0.501</td>
<td>-0.101</td>
<td>0.811</td>
</tr>
<tr>
<td></td>
<td>Is the correlation significant?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>R squared</td>
<td>0.632</td>
<td>0.624</td>
<td>0.654</td>
<td>0.613</td>
<td>0.112</td>
<td>0.251</td>
<td>0.010</td>
<td>0.657</td>
</tr>
<tr>
<td>ICODE</td>
<td>Pearson r</td>
<td>0.380</td>
<td>0.404</td>
<td>0.395</td>
<td>0.344</td>
<td>0.426</td>
<td>0.212</td>
<td>-0.080</td>
<td>0.372</td>
</tr>
<tr>
<td></td>
<td>Is the correlation significant?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R squared</td>
<td>0.144</td>
<td>0.164</td>
<td>0.156</td>
<td>0.118</td>
<td>0.181</td>
<td>0.045</td>
<td>0.006</td>
<td>0.138</td>
</tr>
<tr>
<td>SHABGARD</td>
<td>Pearson r</td>
<td>0.219</td>
<td>0.256</td>
<td>0.261</td>
<td>0.169</td>
<td>0.362</td>
<td>0.227</td>
<td>0.255</td>
<td>0.254</td>
</tr>
<tr>
<td></td>
<td>Is the correlation significant?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R squared</td>
<td>0.048</td>
<td>0.065</td>
<td>0.068</td>
<td>0.029</td>
<td>0.131</td>
<td>0.052</td>
<td>0.065</td>
<td>0.065</td>
</tr>
<tr>
<td>UNPACK</td>
<td>Pearson r</td>
<td>0.718</td>
<td>0.723</td>
<td>0.717</td>
<td>0.679</td>
<td>0.791</td>
<td>0.668</td>
<td>0.222</td>
<td>0.723</td>
</tr>
<tr>
<td></td>
<td>Is the correlation significant?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R squared</td>
<td>0.515</td>
<td>0.523</td>
<td>0.514</td>
<td>0.461</td>
<td>0.626</td>
<td>0.447</td>
<td>0.049</td>
<td>0.481</td>
</tr>
<tr>
<td>XAKEPOK</td>
<td>Pearson r</td>
<td>0.723</td>
<td>0.711</td>
<td>0.631</td>
<td>0.612</td>
<td>0.412</td>
<td>0.502</td>
<td>0.006</td>
<td>0.626</td>
</tr>
<tr>
<td></td>
<td>Is the correlation significant?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R squared</td>
<td>0.522</td>
<td>0.506</td>
<td>0.398</td>
<td>0.375</td>
<td>0.169</td>
<td>0.252</td>
<td>0.000</td>
<td>0.392</td>
</tr>
</tbody>
</table>
Figure 12. **Correlation between Reputation score and Aggregation of Factors (sum of number of posts, total number of replies, total number of views, message length, code length, attachment and tenure).**

**All graphs are plotted on a logarithmic scale to include the large range of values**
SAMPLE QUERIES

Find highest interest posts and topics.

```
SELECT ad.title, postRank, numOfPosts, ad.numOfView
FROM vctoolthread ad, vctoolposts ap, vctoolauthor aa
WHERE ad.threadID=ap.threadID AND ap.authorID=aa.authorID
GROUP BY ad.threadID
ORDER BY ((CAST(postRank AS SIGNED)*5) + CAST(numOfPosts AS SIGNED)) + CAST(numofView AS SIGNED)/3 DESC, LENGTH(numofView) DESC, CAST(numofView AS SIGNED) DESC
```

Find attachment names and rank them by postRank and number of views.

```
SELECT title, flatcontent, subforum, numOfView,
SUBSTRING_INDEX(SUBSTRING_INDEX(contentWithHTMLTag, 'Views: ', -1), 'Size:', 1) Views,
ROUND(SUBSTRING_INDEX(SUBSTRING_INDEX(contentWithHTMLTag, 'Views: ', -1), 'Size:', 1)/REPLACE(numofView, ',', '')*100,0) "%Download/Views",
SUBSTRING_INDEX(SUBSTRING_INDEX(contentWithHTMLTag, 'Name: ', -1), 'Views:', 1) Attachment,
SUBSTRING_INDEX(SUBSTRING_INDEX(contentWithHTMLTag, 'Size: ', -1), '>',' ', 1) Size
FROM vctoolposts vcp, vctoolthread vd
WHERE vcp.threadID=vd.threadID AND contentWithHTMLTag LIKE '%attachment%' AND contentWithHTMLTag LIKE '%Views:%'
ORDER BY CAST(porusrank AS SIGNED) DESC, ORDER BY CAST(VIEWS AS UNSIGNED) DESC
```

Find factors that contribute to reputation score.

```
SELECT distinct P1.authorID as "Author", count(distinct P1.PostID) as "NumofPosts", sum(T1.numOfPosts) as "Total#ofReplies", sum(T1.numOfView) as "Total#ofviews", sum(char_length(P1.flatcontent)) as "MsgLength", sum(char_length(C1.sourcecode)) as "CodeLength", sum(P1.hasAttachment) as "Attachment", datediff(date_format(NOW(), '%Y-%m-%d'), STR_TO_DATE(P1.postDate, '%d %M %Y')) AS "Tenure", CAST(A1.reputationscore as SIGNED) AS "ReputationScore"
FROM anonauthor AS A1, anonthread AS T1, anonposts AS P1
   left outer join anoncode C1 ON C1.postid = P1.postID
WHERE P1.threadID = T1.threadID AND CAST(A1.reputationscore as SIGNED) <> 0
GROUP BY P1.authorID ORDER BY CAST(A1.reputationscore as SIGNED) DESC
```