“WHOIS” Selling All The Pills

Tommy Stallings¹, Brad Wardman, Gary Warner, and Sagar Thapaliya
(1) tds2@uab.edu
University of Alabama at Birmingham

Abstract - Spammers continue to distribute malware, phishing attacks, and counterfeit products to Internet users through emails. The traditional response is to block these emails, but as in other cybercrime fields, law enforcement is realizing the response should be to deter spammers by prosecution. The objective of this research is to enable law enforcement the ability to investigate and analyze related spammed domains in more depth in order to identify trends and potentially key targets that are responsible for creating spam domains. A prototype was developed to examine lists of domains by gathering key components about the information used to register each domain. Additional information on the domain such as the IP address and the Autonomous System Numbering (“ASN”) assignment is also collected. The gathered information serves as input to a clustering algorithm to group seemingly unrelated domains. These clusters are visualized in i2 Analyst Notebook charts that enable law enforcement to quickly target the potential prime suspects in the larger clusters as well as eliminate possible legitimate websites that formed in the smaller clusters. Along with the clustering software that was developed, information was also collected from the UAB Spam Data Mine and analyzed in comparison to the results of the clustering software to reveal a very in-depth pattern of spam domains’ locations across time. These methods demonstrate the effectiveness of automated solution that researchers can provide law enforcement, by quickly analyzing open source intelligence, like the registration information of a website.

Introduction

The Internet has opened up a path to the free exchange of information and ideas. This open architecture has helped to spur increases in productivity and knowledge sharing throughout the world. However, the easy accessibility, as well as lack of oversight on the web, has created a digital haven for various forms of criminal activity. Crimes that are consistently punished in the physical world such as bank robberies, identity theft, illegal pharmaceutical sales, or distributing counterfeit goods, go relatively unchecked in the digital realm.

“Criminals can raid bank accounts without even leaving home,” he said. “Purveyors and
consumers of child pornography have alarmingly turned to computers to conduct their business. Worse yet, children can be preyed upon in our very own homes with a few clicks of a mouse." (Campbell, 2004).

Often times the problem does not arise from deciding whether or not a crime has been committed, but in finding an effective method that quickly and accurately determines the scope and magnitude of the perpetrator's organization or a criminal's nefarious activities (Wardman, Warner, McCalley, Turner, & Skjellum, 2010). “Gathering and analyzing evidence in cases of cybercrime is a crucial problem to be solved. It requires not only special tactics of investigative and organizational actions, but also particular knowledge of computer hardware and software.” (Golubev, n.d.).

Law enforcement currently identifies several prevalent types of online crimes, but these crimes are considered difficult to prosecute effectively. One such cybercrime is counterfeit banking websites, also known as phishing. The principle behind this scam is to deceive unwary Internet users into thinking the website they are visiting is owned by an organization to which the users are affiliated (Wardman & Warner, 2008). Often, the scam will convince the user that some form of unusual activity has taken place on their account and the only way to fix the error is for them to verify their user credentials and private information (Jakobsson & Myers, 2006). Instead of fixing the problem, the user’s sensitive information is sent to a criminal, where it is used for a variety of purposes such as withdrawing money (Li & Schmitz, 2009) or selling the information in Internet chat rooms (Jakobsson & Myers, 2006). Another form of cybercrime is the distribution of malware through the execution of code on a malicious website. This technique is often referred to as a “drive-by” (Provos, Mavrommatis, Rajab, & Monrose, 2008). Malware is malicious software (i.e., viruses, trojan horses, and worms) that is used to provide unauthorized access to a computer system, such as allowing the installation of back doors into the machines, stealing of personal information, and many other problems. Often, the user is not even aware that their system has become compromised.

Spammed emails are often the delivery mechanism for such attacks. Spam is unsolicited email messages sent in bulk, typically offering counterfeit or non-existent goods such as watches or various ‘enhancement’ pills (Cranor & LaMacchia, 1998). These websites are usually illegal in the U.S. because their offered services fall under a violation of a federal law, such as the Canned Spam Act (Federal Trade Commission, 2009), or because selling controlled substances without a pharmaceutical license and valid prescription is illegal (U.S. Department of Justice). Some of these sites also use the acquired credit card information for other illegal activity.

While creating or running such websites is illegal, a lack of consistent prosecution and punishment does little to deter the criminals. There are several methods that are currently employed when dealing with illegal websites, however these methods are ineffective. One common method is to take down the website. While the content may be removed, it does not send a message that this type of action is not tolerated. Another solution for targeting these sites is the ‘whack-a-mole’ method, which involves simply picking certain websites at random and attempting to bring them to justice. While this does send a message to criminals that there is a chance of prosecution, it is ‘hit and miss’ and does not place a priority on the bigger criminals online.

In both physical and cybercrimes, it is important to effectively apply resources to match the severity of the crime. Due to the complexity and relative anonymity of the Internet, it remains difficult for investigators to determine without time consuming efforts, the prevalence of a particular criminal or scheme. For example, a cybercriminal might run a phishing scam that only takes $5 from each person’s bank account. The loss of five dollars to one person is minor and would not be worth an investigator's time, however if the criminal stole five dollars from 100,000
people it becomes a very serious offense worth prosecuting. Therefore, it is imperative to utilize valuable resources to organize online crimes by frequency and severity. Investigators should target prevalent cybercriminals rather than minor offenders. The state of Virginia has already begun to crack down on cybercrimes, “Besides child pornography and Internet fraud the force will concentrate on illegal “spamming”, computer hacking and intrusion, and intellectual property crimes.” (Campbell, 2004).

This research can help address the limited investigative resources by eliciting prevalent patterns in spammed domains as well as providing a much faster approach to identifying prime suspects. In this research, a case study is conducted on domains mainly used for hosting online pharmacies that sell controlled substances without a prescription or a pharmaceutical license to do so. A prototype was developed to gather WHOIS and domain information and cluster the retrieved data in order to determine the prominent criminals.

**Related Work**

Many methods have been employed in the efforts to detect, prevent, and trace spam as well as map out spam campaign behavior. Some spam investigation and clustering techniques use fuzzy matching to cluster text commonly found in the body or subject line of a spam message (Wei, 2009), or natural language processing techniques to generate feature sets via a weighted voting algorithm (Saeedian & Beigy, 2008). Other researchers have clustered spam using a vector space model generated by a spherical k-means algorithm to find a centroid of a cluster to use as a representative for the cluster. This enables labeling of the cluster’s members as ‘spam’ or ‘not spam’ (Sasaki & Shinnou, 2005).

As the various spam detection techniques have evolved, so too have the spamming techniques. Spammers have begun to employ image-based spam messages in which the put their spam messages and URL links in images and spam the images (RedCondor Inc., 2009). While this technique successfully protects the spam message from text-based detection, image-based detection techniques have proven to be successful in clustering and labeling image-based spam. These techniques detect images that are nearly the same, more formally known as “near duplicate detection” (Mehta, Nangia, Gupta, & Nejdl, 2008) or by image segmentation, where the images are broken down into various components and tested in different ways with other images (Zhang, Chen, Chen, & Warner, 2009).

Although all of these techniques have proven to be effective defenses against spamming, and for spam clustering, they fail in terms of helping law enforcement officials to easily track down the parties responsible for the spam campaigns and fraudulent websites. Some techniques have been employed in various studies to aid clustering techniques by looking at the IP addresses hosting the domains (Wei, Sprague, Warner, & Skjellum, 2010). Other researchers have resorted to a more sophisticated implementation of image-based clustering by first performing content-based clustering. Then the approach fetches the main index page of the spam domains in the bodies of the messages to do the image clustering. This clustering links the content-based clusters together with the visual inspections of the images gathered from the spam domains (Chen & Zhang, 2009). This technique performs well with clustering related spam campaigns together; however, the sender information that was extracted in the content-based process can be easily spoofed which would leave investigators with little information to go on. Other detection and clustering techniques also suffer in this way.

By gathering the WHOIS information on the spam domains we can see who these spam domains really belong to and cluster the domains (and spam) based on that. If there are multiple spam campaigns spamming thousands of different domains in different formats, the campaigns would easily be identified and clustered as separate, unrelated campaigns. However, by collect-
Tommy Stallings, Brad Wardman, Gary Warner, and Sagar Thapaliya

By gathering the WHOIS data on the domains and collecting the ASN information on the IPs, the domains will link together and reveal that different spam campaigns are in fact related, even though they appear to be different. Furthermore, by tracking the dates that the domains were first and last observed, the information that has been collected can be laid out across a time line to determine trends in spamming activity based on the domains being spammed. The implementation of using WHOIS data has been used before, but only in a limited fashion such as to validate relationships of previously clustered data (Wei, Sprague, Warner, & Skjellum, Mining Spam Email to Identify Common Origins for Forensic Application, 2008).

The Data Set

A large increase in volume of .ru (country code for Russia) domains in the UAB Spam Data Mine was detected, prompting further investigation as to why there had been such an increase. Investigators of spamming domains can observe when the domains were created and how they were used to spam by reviewing information stored in the UAB Spam Data Mine. Initial analysis of the domains reveals that the first and last dates of messages sending these domains were spammed on average for 5.2 days and had a median of 1 day. There were 236 domains that have been included in spam over 100 days. Several of these domains were associated with large Russian Internet companies. These domains make up smaller clusters not detailed in this paper. However, the vast majority (97%) of the domains in the data set were spammed for three days or less. Figure 1 illustrates the percentage of all .ru domains spammed over a set number of days. While this information is useful to investigators, it doesn’t help with identifying where the spamming domains are hosted, or who is responsible for creating the domains.
Methodology

To collect information about who is responsible for creating the spam domains and where they are being hosted, "WHOIS" requests are employed to collect basic registration information on the domains. The Russian country code domains (i.e., www.somedomain[.ru]) were parsed from spammed emails observed in the UAB Spam Data Mine and compiled into a list of domains that were going to be targeted for the WHOIS requests. Registrants of domain names are required to register and purchase the domain from a domain registrar such as GoDaddy.com. This registration information is kept by the registrar’s office and publicly hosted on the registrar’s WHOIS server. The WHOIS prototype for this study obtains the registration information, extracts fields, and inserts the parsed fields into a database. While the collection of the WHOIS information sounds simplistic, there are a handful of challenging obstacles to overcome.

<table>
<thead>
<tr>
<th>Registrant:</th>
</tr>
</thead>
<tbody>
<tr>
<td>GoDaddy.com, Inc.</td>
</tr>
<tr>
<td>14455 N Hayden Rd #226</td>
</tr>
<tr>
<td>Scottsdale, Arizona 85260</td>
</tr>
<tr>
<td>United States</td>
</tr>
</tbody>
</table>

| Domain Name: **GODADDY.COM** |
| Created on: 02-Mar-99 |
| Expires on: 02-Mar-19 |
| Last Updated on: 08-Dec-09 |

<table>
<thead>
<tr>
<th>Administrative Contact:</th>
</tr>
</thead>
<tbody>
<tr>
<td>GoDaddy.com, Inc., GoDaddy.com, Inc. <a href="mailto:dns@jomax.net">dns@jomax.net</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Technical Contact:</th>
</tr>
</thead>
<tbody>
<tr>
<td>GoDaddy.com, Inc., GoDaddy.com, Inc. <a href="mailto:dns@jomax.net">dns@jomax.net</a></td>
</tr>
<tr>
<td>GoDaddy.com, Inc.</td>
</tr>
<tr>
<td>14455 N Hayden Rd #226</td>
</tr>
<tr>
<td>Scottsdale, Arizona 85260</td>
</tr>
<tr>
<td>United States</td>
</tr>
<tr>
<td>4805058800 Fax -- 4805058844</td>
</tr>
</tbody>
</table>

| Domain servers in listed order: |
| CNS1.SECURESERVER.NET |
| CNS2.SECURESERVER.NET |
| CNS3.SECURESERVER.NET |

| Domain name: enom.com |
| Administrative Contact: |
| eNom, Inc. |
| DNS Manager (domains@demandmedia.com) |
| +1.4259744689 |
| Fax: +1.4259744791 |
| P.O. Box 7449 |
| 15801 NE 24th Street |
| Bellevue, WA 98008 |
| US |

<table>
<thead>
<tr>
<th>Technical Contact:</th>
</tr>
</thead>
<tbody>
<tr>
<td>eNom, Inc.</td>
</tr>
<tr>
<td>DNS Manager (<a href="mailto:domains@demandmedia.com">domains@demandmedia.com</a>)</td>
</tr>
<tr>
<td>+1.4259744689</td>
</tr>
<tr>
<td>Fax: +1.4259744791</td>
</tr>
<tr>
<td>P.O. Box 7449</td>
</tr>
<tr>
<td>15801 NE 24th Street</td>
</tr>
<tr>
<td>Bellevue, WA 98008</td>
</tr>
<tr>
<td>US</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Registrant Contact:</th>
</tr>
</thead>
<tbody>
<tr>
<td>eNom, Inc.</td>
</tr>
<tr>
<td>DNS Manager ()</td>
</tr>
</tbody>
</table>

| Fax: |
| P.O. Box 7449 |
| 15801 NE 24th Street |
| Bellevue, WA 98008 |
| US |

| Status: Locked |

| Name Servers: |
| Dns11.enom.com |
| Dns12.enom.com |
| Dns13.enom.com |

| Creation date: 24 Oct 1997 00:00:00 |
| Expiration date: 21 Sep 2013 17:45:00 |

Table 1: The left column is WHOIS registration information from GoDaddy while the right column is registration information from eNom.
One major problem encountered with parsing registration information is the lack of standard format for registration information. A key component of the WHOIS prototype is the use of regular expressions to find the fields in the information. Many registrars have similar formatting, but slight variations that would render regular expressions ineffective. For example, the registration information format for GoDaddy can be viewed in the left column of Table 1, while the registration information format for eNom can be viewed in the right column. Both registrars have similar structure; however, there are clearly differences in the two structures that make it difficult to parse using generic regular expression. An example of such a case is the creation date of these websites. The words before the date read “Creation Date:” and “Created on:” for both columns respectively, so a regular expression looking for the word “Creat” followed by more letters and possible spaces may prove to be too unreliable. A different attempt may be to try to use the format of the date for a regular expression but again this is not reliable either seeing as there are many different formats for dates. The eNom registrar’s information in the right column includes the time as well as the date. This further demonstrates the differences in information formatting.

The only other solution is to create a parser for each registrar’s specific format; however, there are a limitless number of different registrars for .com domains. This gets even worse when you need the information for domains that end in country codes such as “.cn” or “.ru.” ICANN allows individual countries the ability to have control over their own country code domains (e.g., .cn, .ru, or .us). To retrieve the registration information for a .cn domain, the network information center responsible for keeping registration data for that country code must be contacted. The Chinese country code (.cn) domain information or a referral for more information on the website located somewhere else, can be found at Asian Pacific Network Information Center (APNIC). For registration information on the .cn domain, a query would be submitted to the APNIC WHOIS server and the registration information about the website would be returned. However, the information returned is not only in English, but also Chinese. Even though English text is present, the text is more of a translation and does not match the common English key words that would be sought for in the information during parsing.

The most reliable approach to gather WHOIS information as well as additional domain information is to parse lexically identifiable strings such as email addresses and name servers. An example of using regular expressions is finding the standard format of an email address using:

```
[\d\w!#$%&\'\(\)\*\+\-,\./;=\?\^\`\{\|\}~\]\+@\S+\.(\w{2,6})
```

This regular expression is searching for one or more digits, letters, and all of the special symbols allowed in proper email address format, followed by an “@” symbol, followed by any other potential non-whitespace characters for the domain that the email address belongs to, followed by a period, followed by 2 to 6 letters. Email addresses always have this type of format and therefore, they can always be caught like this. Name servers, which are clearly listed in the registration information examples in table 1, follow the same principle, except that they are similar to domain names. To ensure a domain name is not parsed out, a check to the parse statement ignores strings containing “www.” as it is probably the website’s domain name.

The email address is useful to investigators as the domain cannot be activated until it has been verified via an email message. If a criminal wanted to create thirty fraudulent domains, then he would have his email address listed on each domains’ registration information. We can therefore use that email address to cluster websites. Skillful criminals could use thirty different email addresses to activate the thirty domains. This is where the use of other domain information such as the name servers could be applied.
Since some criminals register many domains with the same registrar, they can sometimes be clustered by sharing the same name servers. This clustering can occur because when the criminal does not explicitly state the name server they will be using, the registrar will provide the domain with one. When large clusters of suspicious domains appear, a quick investigation of the email addresses may reveal that the domains are owned by the same individual because that individual may use those email addresses for other things which would link them together inferring that they are either being used by the same person, or they belong to people that work together in a group; moreover, clustering based on name servers can infer relationships in how the domains are configured. Registrars also have the ability to assign domains with an IP address, as it has a license to sublease a specific quantity of IP addresses to domains based on what class of IP license that the registrar owns. This license is identified by the Autonomous System Numbering information. When a domain name is entered into the address bar of a web browser, it queries name servers for the domain it is searching for, and in return, gets the IP address for that domain. The browser then follows referrals until it reaches the name server that the IP is hosted on. Having the name server information is very useful, but there is still a slight problem with obtaining the name server from the WHOIS information. Even though parsing the name server is reliable, the standards for keeping registration information are still a wild card. Not every registrar lists the name server for that domain in its registration information. For this reason, other information is collected to ensure that there is enough data for reliable leads for investigation.

More features of the domain that are collected are the associated IP address and ASN information. The solution to obtaining each website’s IP is to ping the website, capture the returned information, and parse the IP address. The IP address is used to look up the ASN using a built-in Unix WHOIS command instructed to query the WHOIS server at Team Cymru (Team Cymru). This WHOIS server that keeps track of all ASN assignees. The returned information from the query is parsed and the ASN name and number are extracted.

An agglomerative hierarchical clustering algorithm was developed to cluster the domains based on the collected data mentioned above. The clustering algorithm initially joins domains based on each feature. For example, when two domains are hosted on the same IP address, those domains will be grouped together in the same cluster. Once all the domains are clustered with respect to each feature, the feature clusters are then joined based on the domains. The results of this joining of information are analyzed below.

Discussion

The discussion of this paper is broken into three sections: spam clustering analysis, a time-based analysis, and finally, the results of a custom developed visualization tool. Analysis of the results in the spam clustering presented problems with respect to time as the clustering process provided a high level view of how the spam was being registered, however, it was not clear how it evolved over time. Therefore a section was added that provides insight into how .ru spam domains were being created and reused. However, the question remained of which entities were responsible for the domains over a certain time period. In response to this question, a custom visualization tool was developed for visualizing spam and WHOIS data. The usefulness of this tool is described and accompanied by a discussion of the visualized data. Finally, summarization of potential findings is presented.

Spam Clustering Analysis

To demonstrate the effectiveness of the registration information analysis, 10,808 Russian pill spam domains were inserted into a list and were handed off to the WHOIS prototype. The WHOIS prototype took around six hours to collect the registration information as some WHOIS servers required the program to throttle its queries due to bandwidth restrictions. The clustering of the
information revealed twelve email addresses that were major contributors for registering the illegal pill websites. Each email address inserted into the database was tagged with a unique domain id number, and that domain id number is a unique identifier for the domain’s registration information. Furthermore, an actual copy of the spam message containing that particular domain name can be extracted from the UAB Spam Data Mine for more evidence. Prominent email addresses are generally the primary focus for investigators; however, the investigators can also use Autonomous System Numbering assignees to potentially understand where these websites are being hosted and who is hosting the websites. The ASN entities distribute IP addresses from their assigned IP range to the domain registrars. These registrars can provide the domain with an IP address, as well as to ISPs that grant individual IP addresses to the public. The most-frequently occurring ASN out of approximately eleven thousand domains was a Korean ASN block called KRNIC-ASBLOCK-AP. This ASN block is comprised of sub-groups of Korean ASNs but the ASNs are grouped together in a “block”. This block is associated with 2,970 total websites hosted on its IP address range, accounting for about one-third of the domains in the list. The second largest ASN was a British ASN called BESTISP-AS which had a total of 2,680 websites hosted on its IP address range. If just these two ASNs were to revoke the pill spam sites’ IP addresses, then the total count for this tremendous cluster would be cut in half. The top 10 ASNs are listed in Table 3.

Table 2: The top twelve email addresses and the number of domains registere

<table>
<thead>
<tr>
<th>Number of Registered Domains</th>
<th>Email Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>1594</td>
<td><a href="mailto:voycehovskiy68@mail.ru">voycehovskiy68@mail.ru</a></td>
</tr>
<tr>
<td>515</td>
<td><a href="mailto:altsrv@googlemail.com">altsrv@googlemail.com</a></td>
</tr>
<tr>
<td>407</td>
<td><a href="mailto:wasa1974@hotnail.com">wasa1974@hotnail.com</a></td>
</tr>
<tr>
<td>403</td>
<td><a href="mailto:maxbaev@invitra.ru">maxbaev@invitra.ru</a></td>
</tr>
<tr>
<td>352</td>
<td><a href="mailto:skrsky@anyweb.info">skrsky@anyweb.info</a></td>
</tr>
<tr>
<td>302</td>
<td><a href="mailto:mikepanin1990@gmail.com">mikepanin1990@gmail.com</a></td>
</tr>
<tr>
<td>284</td>
<td><a href="mailto:annatrsv@gmail.com">annatrsv@gmail.com</a></td>
</tr>
<tr>
<td>220</td>
<td><a href="mailto:wasa1974@hotnail.com">wasa1974@hotnail.com</a></td>
</tr>
<tr>
<td>170</td>
<td><a href="mailto:i-nikitin.2000@gmail.com">i-nikitin.2000@gmail.com</a></td>
</tr>
<tr>
<td>125</td>
<td><a href="mailto:jeez@gmail.com">jeez@gmail.com</a></td>
</tr>
<tr>
<td>119</td>
<td><a href="mailto:altsrv@gmail.com">altsrv@gmail.com</a></td>
</tr>
<tr>
<td>105</td>
<td><a href="mailto:nikolatesla@mails.ru">nikolatesla@mails.ru</a></td>
</tr>
</tbody>
</table>

The last part of the analysis was the name servers which provided unexpected results. For example, approximately 3,000 websites are hosted on the KRNIC’s IP address range, which are Korean ASNs; yet all of the name servers are located in China. The second largest ASN, BESTISP, is British, but only around 55% of the websites in its IP range had name servers hosted in the UK. Further analysis of 2,680 websites hosted on BESTISP revealed that roughly 1,350 associated name servers are hosted in Netherlands. The other 4% are either hosted in China or Russia, while around 1% resides in Czech Republic. Most of the remaining name servers for all other ASNs are hosted in either Russia or China; however, there are still a number of name servers hosted in the UK, Malaysia, and Netherlands. This location pattern is interesting because it is leading to one of two possible hypotheses about a trend emerging from analyzing the clusters. However, to fully understand this trend, sections of the i2 chart will be examined. Appendix A contains the entire main cluster which consists of 9,488 of the 10,808 domains analyzed in this research.

The color code of the clustered information is as follows - green links are connections from the website to the ASN that the IP resolved to. Blue links are the connections from the website to the name server(s) listed in the website’s registration information. Red links are connections.
from the website to the email address that was used to register the website. And finally, black links are the connections from the website to its IP address.

Analysis of the largest wedges of each color illustrates the top investigative elements in this research. These elements of interest are identified visually by the points of inflection that are created by having so many connections. Starting at the very top of the chart, shown in Figure 2, is the KRNIC ASN Block. Below the large portion of green lines are associated domains. Each row of domains will refer to a specific tier in the cluster with this first occurrence of domains signifying Tier One. These domains then connect to a manageable number of name servers and email addresses which are the middle inflection points just below Tier One. Further analysis of the middle inflection points reveals the email address mikepanin1990@gmail.com which is identified by this research as one of the twelve most common email addresses used to register .ru spam domains. This email address is accountable for a noticeable portion of the connections between the domains of Tiers One and Two because it is associated to many domains hosted on KRNIC and a Canadian ASN, MTNCABLE. The other major points of inflection are the Chinese name servers which are hosting sites from KRNIC and MTNCABLE. The Tier Two row indicated in Figure 2 is composed of domains which are connected to both Tiers One and Three by common ASN or name server. One of the main attributes that forms the bridge-like link through Tier Two is the Colohost ASN which is the main inflection point located below Tier Two in Figure 2. Adjacent to MTNCABLE is another top 12 email address, altsrv@googlemail.com. This email address accounts for 515 of the websites in MTNCABLE which is more than half of the sites in MTNCABLE’s IP range. MTNCABLE is the third largest ASN in this study with 948 sites within its IP range, yet all of its sites are on name servers in China. This elicits yet another unexpected result returned from the geo-coding of the domains’ name servers.
Figure 3 shows Tiers Three and Four of the cluster which are tied to Tier One mainly by the ASNs BESTISP and COLOHOST and the email address mikepanin1990@gmail.com. BESTISP is the British ASN described previously that uses name servers either in the UK or the Netherlands. A significant portion of these domains follow the unusual practice of using their own domain for the name server resolution. As an example, the domain refilluther29s.ru has a name server called ns1.refilluther29s.ru. In addition, most legitimate domains are hosted on at least two name servers. The domain name choices are also suspicious, as if created by selecting random words from a category list composed of a word related to pharmaceuticals, such as drugs or med, followed by a person's first or last name followed by two numbers and a letter. For example, “ns1.drugstodd48m.ru” and “tabtom34o.ru” are examples of many patterns observed.

Looking at the red wedges in Figure 3, three more of the top email addresses in this tier are easily observed, especially the most commonly observed email address, voycehovskiy68@mail.ru used to register about one-third of the websites in this tier. The other two email addresses are prominent links from Tier Three to Tier Four, along with a couple of Chinese and Russian name servers.
Figure 4 displays the fourth tier down to the ninth tier at the bottom of the pill site cluster. The very top of Tier Four is predominately connected by two of the top 10 ASNs CHINANET-BACKBONE and Eveloz. The domains connected to Eveloz, an ASN in Panama, are using name servers hosted in the UK, the Czech Republic, Vietnam, or Malaysia, while the domains on IPs in the CHINANET-BACKBONE use Chinese or Russian name servers. This behavior indicates a migration trend between the locations of name servers. Along with the name server geography, Tier Four also contains two of the top 10 ASNs that are the main connections between Tiers Four and Five. From Tier Four to Tier Five the primary connections are name servers based in Russia and China, although there are also name servers in Germany and the United Kingdom. Further down Tier Five there is an emergence of a dominant trend of Russian name servers. There are a few Belgium name servers but no sign of Chinese or British name servers. By comparing the geographical locations of the name servers from Tier One down to the bottom-most tier, a migration pattern emerges. This migration pattern can be the result of two possibilities depending on time.

If the domains in the Russian domain data set from the UAB Spam Data Mine are stretched out across a long period of time, then it is possible that this population growth in name servers in Russia are from international cyber-criminals migrating their illegal online activities to Russia. However, if the time span of these observed dates is only for a short period of time, then this could indicate that there is either a group of Russian criminals working together, or that there is a growing trend of online crime in Russia. Because the spam dates of the Russian domains in the UAB Spam Data Mine are observable, it is possible to do a timeline analysis of these domains.
Time-Based Analysis

Figure 5 contains three different charts representing domain and IP activity for the top ten countries that were observed to be hosting the websites in the data set. The chart in the upper left corner is the number of .ru domains per country per month over the course of a year, whereas, the chart in the upper right corner shows the number of unique IP addresses observed from the top ten countries per month over the same time period. Finally, the chart on the bottom shows the number of unique Russian domains new to the UAB Spam Data Mine spammed per day over the course of a year. This chart demonstrates exactly what days a new spam campaign (sending of a domain) started. The three charts in Figure 5 not only show the timeline for the activity, but also the emergence of a spam campaign. It is observed that in the beginning of the year the number of newly created .ru domains being spammed was very little, which is indicative of the same domains being heavily spammed. After half of the year has passed by with the same domains being heavily spammed, it is easy for the domains to get blacklisted which in turn effectively cripples the domain from being spammed. It is believed that as spammers became aware of the blacklisting being implemented by spam filters, they began to play with hosting options.

Figure 5 – Domain and IP activity for .ru top level domains
By creating many domains with only a few different IP addresses, they were able to out run the domain name blacklisting but only for a month or so until the IP was blacklisted. The spammers also created many domains with many IP addresses which was able to defeat the blacklisting. The large spikes in all three charts help with the perspective of this massive domain/IP explosion. Spammers created more domains in just a couple of days than they had in six months.

Description of the Visualization Tool

Considering the big volume of WHOIS data and thus the challenge it possesses to get insight into the information it contains, software was developed to visualize the data for a closer look. As input, the software took the spam and WHOIS information including the spammed domains, first and last dates of email messages sending these domains, IP address of the domains and email address associated with them. Then an interactive bar diagram was developed to visualize this data. The tool provides an interactive interface with color codes to allow easy exploration of the data.

The major goal of this visualization tool was to look at evolution of spam domains associated with a given email address through time. It could be altered to look at the relation of those domains with associated IP address, to see if different domains existing at different points of time shared same IP address. In addition, finding patterns within the creation domains by different email addresses was an interest too.

First the data was reorganized into clusters, with each cluster representing details of the activities of an email address. Using the starting and ending dates for spammed message related to a domain as an event and each cluster included the following information: set of dates when events occurred, total number of domains launched on each event date, total number of domains shut down on each date, and the set of unique IP addresses associated with the cluster.

The capabilities and labels for the visualization’s bar diagram of the email clusters can be found below:

- X-axis – sorted event dates
- +ve Y-axis – the number of new domains started at a given date
- -ve Y-axis – the number of websites still operating/active by the end of a day
- color bands of +ve Y-axis bars – each unique color represents a unique IP address for the cluster
- zoom in and out – provides closer look at desired regions in visualization

Usefulness of the Visualization Tool

The tool provides an effective way to look at the spam and WHOIS data. By looking at the bar diagram produced for an individual registrant (i.e. email address), one can easily see the following information:

- Time span between which the spammer was active
- Size of the attacks made at different dates
- IP addresses involved
- Size and dates of attacks made using each unique IP address (i.e. from different places)
Furthermore, visualizing multiple clusters at the same time allows comparison of activity of two registrants. It can reveal interesting facts like: if two people were active at the same time, followed same pattern, working in same location. If one of the registrants stopped being active and the other started at a single point in time, it may indicate that a single person is involved in both attacks but changed identity and location.

**Analysis of the Visualization Results**

![A demonstration screenshot of the domain plot for wasa1974@honnail.com using the visualization tool.](image)

The bar diagrams presented in Figures 6 and 7 as well as Appendix B, for a larger picture, illustrates the ability of the tool to present a visualization to the user with the information that would be otherwise difficult obtain. Figure 6 is a screenshot of the tool with the email address wasa1974@honnail.com as input. This screenshot demonstrates every utility of the tool except the tools ability to zoom in and out, which is accomplished by holding down the left mouse button and moving the mouse forward or backward. As mentioned in the description of the tool, the x-axis is time, the bars above the x-axis are representative of the number of unique domains spammed that day, and finally the light blue bars below the x-axis are the number of domains still active after the day it was spammed. Each color variation, delineated by the white stripes, represents the number of domains on a particular IP address.

As illustrated by the multicolored bars in Figure 6, from the July 6th to July 17th wasa1974@honnail.com registered many domains on many different IP addresses. Around July 24th, wasa1974@honnail.com started using three different IP addresses as observed in the large span of bars with the singular red color. On August 23rd, the final day of registered domains by that email address, there is a large black bar that indicates that all of the domains, largest peak of Figure 6, was hosted on only one IP address.
Another example of the visualization tool can be observed in Figure 7. This diagram contains the domain and IP pattern for voycehovskiy68@mail.ru, the email address with the largest number of domains registered. This email address hosted a majority of the domains on a single IP. This may be indicative to the domains being hosted on a bullet-proof server (i.e. a hosting company or server in which the content is difficult to get removed from the Internet). The large sets of black bars were all registered between the beginning and end of July. These large set of black bars as well as the large set of multicolored bars in Figure 6 are consistent with what was observed in July of Figure 5. In addition, the domains from voycehovskiy68@mail.ru and wasa1974@honnail.com had similar patterns of starting in the beginning of July and finishing on August 23rd and both were registering a consistently lower number of domains in August than in July. Given the similar activity observed, it is believed that these email addresses are either the same person, or work together in the same group.

Observations of the top eleven email addresses that were used to register domains in the largest cluster showed that after July 17th, 2010 all of the domains per email address were hosted on only one to five IP addresses. Whereas prior to this date, domains were hosted on many IP addresses. These results are similar to what is observed in Figure 5, however, these IP addresses are only associated to email addresses in the largest cluster and are not representative of all the pill spam activity (i.e. smaller clusters that were not fully analyzed).

Discussion of Findings

As the name server hosting migrates to Russia, the question arises whether criminal activity in Russia is increasing, or whether Russian criminals are encountering difficulties hosting their infrastructure in foreign countries. Appendix A illustrates how the cluster takes the shape of a ladder-like structure with the rungs increasing in size as the chart progresses upwards. The first explanation for this progressive-like behavior would be that these websites have been running for a while. It would seem like it must take a lot of time for all of these websites to be created and hosted, but with a standard website template already developed, it leaves the criminal with little to do but to come up with a name for the website and find a registrar to host the website. By physically visiting the sites and looking at the layouts,
the observation was made that there are only a few variations of these sites which indicates that the criminals are indeed using a standard template. To test the idea that these criminals can rapidly upload their websites to the Internet, all that needs to be done is to look at the creation date of the domains which is provided in the registration information for each domain. Additionally, by keeping copies of the registration information, evidence about these websites can be further preserved as well as enable more in-depth examination of the information. However, reviewing the registration records to see the timeline of the websites will not be necessary because this timeline analysis has already been observed in the section that introduced our data set which effectively eliminates the possibility that these websites have been active for a long period of time.

The timeline analysis on the spamming dates and days spammed reveals that the majority of websites are indeed being created rapidly. The creation dates from the websites that have the top 12 email addresses listed in the registration information span from the first of June to mid-August, while previous to these dates much fewer domains were being registered by many email address. This conclusion also supports the other hypothesis that shows the lack of criminal prosecution to these types of crimes and is almost an encouragement to these criminals. The current tendency for handling websites that sell prescription drugs is to simply shut down the website. Shutting down the website does not send a message that would deter the criminal from committing anymore crime. As a result of this, the criminal has low risk but high reward and can setup another website within hours. Without the use of punishment as a deterrent, the cyber-criminal will simply create more sites, and encourage others to join in on the online cyber-crime world. The smaller tiers at the bottom of the chart are probably coming from people who are “testing the waters” for this type of crime. Since the lower tiers are primarily hosted in Russia, the lack of punishment for these types of crimes in Russia or slower shutdowns could be why these tiers are spawning up, not because criminals are migrating to Russian hosting.

**Conclusion**

It is an injustice to only respond to illegal websites by having them removed from the Internet. It takes too long for law enforcement to collect enough necessary information on all spammed domains and analyze the results before the websites are shut down. Furthermore, with the problems mentioned above about registrars, it is easy for criminals to create new email accounts and register new malicious domains. Investigators lack the ability to prove who the prominent criminals to pursue are in order to make the greatest impact on stopping these websites. The standard approach is essentially like a game of darts – luck. With the guidance of this research, new approaches are introduced to help investigators make significant impact to stop criminals effectively, and put them behind bars.

**REFERENCES**


“Whois” Selling all the Pills


Appendix A
Appendix B