E-crime

CS642: Computer Security

Professor Ristenpart

http://www.cs.wisc.edu/~rist/

rist at cs dot wisc dot edu
Trailrunner7 writes

"It is open season on SCADA software right now. Last week, researchers at ReVuln, an Italian security firm, released a video showing off a number of zero-day vulnerabilities in SCADA applications from manufacturers such as Siemens, GE and Schneider Electric. And now a researcher at Exodus Intelligence says he has discovered more than 20 flaws in SCADA packages from some of the same vendors and other manufacturers, all after just a few hours' work."
Spam, phishing, scams

• Spam
  – unsolicited bulk emails
  – 2006: 80% of emails on web, 85 billion messages a day

• Scam spam
  – Nigerian emails (advanced fee fraud / confidence trick)

• Phishing
  – trick users into downloading malware, submitting CC info to attacker, etc.
  – Spear phishing: targeted on individuals (used in high-profile intrusions)
Spanish Prisoner confidence trick

- Late 19\textsuperscript{th} century
- In contact with rich guy in Spanish prison
- Just need a little money to bribe guards, he’ll reward you greatly
Hi Dear,

I am Mrs. Zarina Al-Usman, I have been diagnosed with Esophageal cancer. It has defied all forms of medical treatment, and Right now, I have only about a few months to live and I want you to Distribute my funds worth Twelve Million Five Hundred Thousand US Dollars to charities homes in your country.

I have set aside 40% for you and your family so keep this as a secret to yourself because this will be my last wish.

Yours Truly,

Mrs. Zarina Al-Usman

WebMail    FDV - MG
Faculdade Viçosa
Your mailbox has exceeded its limit. Your webmail is currently running 99.7% of its Quota limit of 100%. You cannot send or receive email until you have updated your webmail account. To update your webmail account, copy the link below and paste in your browser to request for upgrade.


We are sincerely sorry for any inconvenience this might cause you; we tend to serve you better.
Thanks for your co-operation.
Webmail Update Team
Spam

- The frontend (email recipients)
  - Filtering, classification
  - Psychology, usability
- The backend (email generation)
  - Open email relays
  - Botnets
  - Social structure
    - Affiliates
    - Criminal organizations
Botnets

- Botnets:
  - Command and Control (C&C)
  - Zombie hosts (bots)
- C&C type:
  - centralized, peer-to-peer
- Infection vector:
  - spam, random/targeted scanning
- Usage:
  - What they do: spam, DDoS, SEO, traffic generation, ...
How to make money off a botnet?

• Rental
  – “Pay me money, and I’ll let you use my botnet... no questions asked”

• DDoS extortion
  – “Pay me or I take your legitimate business off web”

• Bulk traffic selling
  – “Pay me to direct bots to websites to boost visit counts”

• Click fraud, SEO
  – “Simulate clicks on advertised links to generate revenue”
    – Cloaking, link farms, etc.

• Theft of monetizable data (eg., financial accounts)

• Data ransom
  – “I’ve encrypted your harddrive, now pay me money to unencrypt it”

• Advertise products
How to make money off financial credentials?

• Money mules
  – Deposits into mules’ account from the victim’s
  – Mule purchases items using stolen CCN, sells them online
  – Mule withdraws cash from ATMs using victim credentials

• Wires money to (frequently) former Soviet Union
Dear Student,

I would like to offer you a new interesting and respectable job! We are looking for people to work as professional distance-based typists. No experience is needed! If you’re eager to use your skills to make some additional cash, then you might want to consider a home typing position!

All data entry operators work from home and are independent contractors. You typically set your own hours and work from home on projects that are enjoyable! Average monthly earnings start from $1000 to $3000 or more.

Requirements:
- Computer with Internet access.
- Good Typing Skills.
- Basic Internet knowledge.
- Basic Computer and Typing Skills.

You will not have to devote full time hours. These assignments can be done on your time. They may be done in Internet cafes or wherever you can get Internet access!

If you are interested just reply to my email!

Best Regards,

Richard Hill
Local Recruitment Manager
Underground forums

<table>
<thead>
<tr>
<th>Category</th>
<th>Threads</th>
<th>Users</th>
<th>Top Subcategory</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>payments</td>
<td>8,507</td>
<td>8,092</td>
<td>paysafecard</td>
</tr>
<tr>
<td>game-related</td>
<td>2,379</td>
<td>2,584</td>
<td>steam</td>
</tr>
<tr>
<td>accounts</td>
<td>2,119</td>
<td>2,067</td>
<td>rapidshare</td>
</tr>
<tr>
<td>credit cards</td>
<td>996</td>
<td>1,160</td>
<td>unspecified cc</td>
</tr>
<tr>
<td>software/keys</td>
<td>729</td>
<td>1,410</td>
<td>key/serial</td>
</tr>
<tr>
<td>fraud tools</td>
<td>652</td>
<td>1,155</td>
<td>socks</td>
</tr>
<tr>
<td>tutorials/guides</td>
<td>950</td>
<td>537</td>
<td>tutorials</td>
</tr>
<tr>
<td>mail/drop srvs</td>
<td>751</td>
<td>681</td>
<td>packstation</td>
</tr>
<tr>
<td>merchandise</td>
<td>493</td>
<td>721</td>
<td>ipod</td>
</tr>
<tr>
<td>services</td>
<td>266</td>
<td>916</td>
<td>carder</td>
</tr>
<tr>
<td></td>
<td>1,539</td>
<td>1,409</td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Top 10 most commonly traded merchandise categories on LC.

Agobot (circa 2002)

• IRC botnet
• Rich feature set:
  – Well-documented, modular codebase
  – IRC-based C&C system
  – Large catalogue of remote exploits
  – Limited code obfuscation and anti-disassembly techniques
  – Built-in data collection
  – Mechanisms to disable antivirus
  – Large set of bot commands
Storm botnet

- Sept 2007
  - Media: 1 – 50 million bots
  - More likely: 10,000s to 100,000s

Features:

- Uses P2P (Overnet/Kademlia)
- Uses fast-flux DNS for hosting on named sites
- Binary has gone through many revisions
- Features of P2P network have evolved with time
- Hides on machine with rootkit technology

Enright 2007
The blue peers count is all peers being probed at a time. This includes live, active, dead, and unknown states. The peers line is not the size of the network. The active line is much closer to the instantaneous size of the network.

It can be seen in the month and year chart that Microsoft made a measurable dent in the network with the MRT Storm (Nuwar) release.
These techniques may already account for wide discrepancies in the estimated size of various botnets seen in the media. With so many groups taking uncoordinated actions with noticeable effects, it is only a matter of time before problems occur. For example, one possible problem would be the effect of a researcher inflating the perceived size of a botnet that is the subject of a criminal investigation. If such a case resulted in a successful prosecution, and a damage estimate were to be derived based on the inflated count of “infected” hosts, multiplied by some estimated cost of cleanup accepted by the court, the resulting damages would be similarly inflated. This is not out of the question, as several cases in the past few years have included evidence obtained by law enforcement agents as to the number of bots under the control of the suspect. It is likely that some of these suspects, even if they admit to the numbers stated, may not know precisely how many hosts they truly did compromise and control.

One final interesting observation, which we have not seen noted in any other research to date, are the downward spikes in the bottom line—the reachable and responsive peers—of the botnet. Geolocating bots enumerated for Naguche botnet
Dittrich and Dietrich, “Discovery Techniques for P2P Botnets”
<table>
<thead>
<tr>
<th>Technique</th>
<th>Description</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monitor endpoint</td>
<td>monitor traffic of a bot</td>
<td>simple, generally applicable</td>
<td>limited view, encryption</td>
</tr>
<tr>
<td>Internet telescopes</td>
<td>monitor random-scan infection attempts</td>
<td>botnet-wide view</td>
<td>limited applicability</td>
</tr>
<tr>
<td>Monitor IRC</td>
<td>record IRC C&amp;C traffic</td>
<td>simple, botnet-wide view</td>
<td>only IRC botnets</td>
</tr>
<tr>
<td>DNS redirect</td>
<td>hijack C&amp;C via DNS</td>
<td>measure infection size</td>
<td>limited applicability</td>
</tr>
<tr>
<td>Sybil monitoring</td>
<td>monitor numerous bots</td>
<td>simple, passive</td>
<td>resource-intensive, limited view, structed P2P</td>
</tr>
<tr>
<td>Botnet crawling</td>
<td>crawl botnet overlay</td>
<td>enumerate large portion of botnet</td>
<td>detectable</td>
</tr>
<tr>
<td>DNS cache probing</td>
<td>probe DNS caches for botnet C&amp;C</td>
<td>simple, passive</td>
<td>loose lower-bound</td>
</tr>
<tr>
<td>DNSBL counter-intelligence</td>
<td>sniff DNSBL traffic, heuristically identify bots</td>
<td>passive</td>
<td>limited applicability</td>
</tr>
<tr>
<td>Flow analysis</td>
<td>detect botnets via flow-based anomaly detection</td>
<td>wide-scale, handles encryption</td>
<td>tailored to IRC botnets</td>
</tr>
</tbody>
</table>
Size estimates from literature as of 2008

<table>
<thead>
<tr>
<th>Study</th>
<th>Method(s) used</th>
<th>C&amp;C’s observed</th>
<th>Largest botnet size infection</th>
<th>Largest botnet size effective</th>
<th>Total # of infected hosts</th>
</tr>
</thead>
<tbody>
<tr>
<td>[13]</td>
<td>IRC monitoring</td>
<td>∼100</td>
<td>226,585</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>[8]</td>
<td>IRC monitoring</td>
<td>∼180</td>
<td>∼50,000</td>
<td>–</td>
<td>∼300,000</td>
</tr>
<tr>
<td>[22]</td>
<td>DNS cache probing</td>
<td>65</td>
<td>–</td>
<td>∼3,000</td>
<td>85,000</td>
</tr>
<tr>
<td></td>
<td>IRC monitoring</td>
<td>&gt;100</td>
<td>&gt;15,000</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>[23]</td>
<td>DNS cache probing</td>
<td>100</td>
<td>–</td>
<td>–</td>
<td>88,000</td>
</tr>
<tr>
<td></td>
<td>IRC monitoring</td>
<td>472</td>
<td>∼100,000</td>
<td>&gt;10,000</td>
<td>426,279</td>
</tr>
<tr>
<td>[5]</td>
<td>DNS redirection</td>
<td>∼50</td>
<td>&gt;350,000</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>[15]</td>
<td>flow analysis</td>
<td>∼376</td>
<td>–</td>
<td>–</td>
<td>∼6,000,000</td>
</tr>
<tr>
<td>[7]</td>
<td>botnet crawling</td>
<td>1</td>
<td>∼160,000</td>
<td>∼44,000</td>
<td>–</td>
</tr>
</tbody>
</table>

Figure 2: Size estimates from the literature. All sizes are the maximum ones given in the appropriate study and the final column represents the total number of infected hosts over all botnets encountered.
Botnet takeover studies

• Spamalytics (Kanich et al., 2008)
  – Storm botnet
  – Rewrote spam to redirect to researcher-controlled websites
  – **Goal:** click-through rate measurement

• Torpig C&C sinkholing (Stone-gross et al., 2009)
  – Torpig botnet
  – Setup researcher controlled C&C server (DNS fastflux)
  – **Goal:** analysis of stolen data
Kanich et al., Spamalytics: An Empirical Analysis of Spam Marketing Conversion, 2008
The victims

Figure 9: Geographic locations of the hosts that “convert” on spam: the 541 hosts that execute the emulated self-propagation program (light grey), and the 28 hosts that visit the purchase page of the emulated pharmacy site (black).

Kanich et al., Spamalytics: An Empirical Analysis of Spam Marketing Conversion, 2008
Observed Conversion Rate

• 350 million email messages delivered
• 26 day campaign
• 28 “sales”
  – 0.00001%
  – 27 of these male-enhancement products
• Statistical significance?
Botnet takeover studies

• Spamalytics (Kanich et al., 2008)
  – Storm botnet
  – Rewrote spam to redirect to researcher-controlled websites
  – **Goal:** click-through rate measurement

• Torpig C&C sinkholing (Stone-gross et al., 2009)
  – Torpig botnet
  – Setup researcher controlled C&C server (DNS fastflux)
  – **Goal:** analysis of stolen data
In fact, the injected content carefully reproduces the user's browser to this content typically consists of an HTML URL from the injection server and injects the returned content into visits the trigger page. The Torpig requests the injection of times it can be launched. The second step occurs when the user attack to whether the attack is active and the maximum number of times it can be launched. The URL the injection server that contains the phishing content to the page where the attack should be triggered we call this page the server. A banking website, Torpig issues a request to an injection server in one of the domains specified in the configuration file. These phishing attacks are very difficult to detect, even for modern browsers. They are designed to automate the decryption. The 'r' server can reply to a request from the client with a reply to the injection server. The server can also send a configuration file to the bot, which is obfuscated using a simple XOR encryption. We call this reply an obfuscation mechanism based on XORing the clear text with a secret key. The 'r' server can send configuration information from its victims, which otherwise may not be observable by security researchers at the end of 411.x and tools are available to automate the decryption. The 'r' server can reply to a request from the client with a reply to the injection server. The server can also send a configuration file to the bot, which is obfuscated using a simple XOR encryption. We call this reply an obfuscation mechanism based on XORing the clear text with a secret key. The 'r' server can send configuration information from its victims, which otherwise may not be observable by security researchers at the end of 411.x and tools are available to automate the decryption. The 'r' server can reply to a request from the client with a reply to the injection server. The server can also send a configuration file to the bot, which is obfuscated using a simple XOR encryption. We call this reply an obfuscation mechanism based on XORing the clear text with a secret key. The 'r' server can send configuration information from its victims, which otherwise may not be observable by security researchers at the end of 411.x and tools are available to automate the decryption.
Stone-Gross et al., Your Botnet is My Botnet: Analysis of a Botnet Takeover, 2009
Figure 3. Unique bot IDs and IP addresses per hour. The number of unique IP addresses per hour provides a good estimation of Torpig’s live population.

Stone-Gross et al., Your Botnet is My Botnet: Analysis of a Botnet Takeover, 2009
Table 1. Data items sent to our C&C server by Torpig bots.

<table>
<thead>
<tr>
<th>Data type</th>
<th>Data items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form data</td>
<td>11,966,532</td>
</tr>
<tr>
<td>Email</td>
<td>1,235,122</td>
</tr>
<tr>
<td>Windows password</td>
<td>1,258,862</td>
</tr>
<tr>
<td>POP account</td>
<td>411,039</td>
</tr>
<tr>
<td>HTTP account</td>
<td>415,206</td>
</tr>
<tr>
<td>SMTP account</td>
<td>100,472</td>
</tr>
<tr>
<td>Mailbox account</td>
<td>54,090</td>
</tr>
<tr>
<td>FTP account</td>
<td>12,307</td>
</tr>
</tbody>
</table>
A credit card validation heuristic that includes the Luhn algorithm was used to validate the credentials from different sources. Number because Torpig uses different data formats to upload stolen credentials from different sources.

It was possible to infer that the compromised accounts were obtained from the password manager of browsers rather than by intercepting an actual login session. Of particular interest is the case of a credit card validation heuristic that includes the Luhn algorithm.

Credential theft was a significant problem, with 47% of the credentials stolen by Torpig and sent to our server in ten days. Torpig observed all of the interested parties, indicating that the infection was large-scale.

The large number of institutions that had been breached made no significant difference. Only a handful of compromised accounts were obtained from the password manager of browsers, rather than by intercepting an actual login session.

On the other end of the spectrum, a large number of companies had been targeted. The top targeted institutions were PayPal, Capital One, RevShare, and HSBC. The top targeted institutions were PayPal, Capital One, RevShare, and HSBC.

The table below summarizes the number of accounts at financial institutions stolen by Torpig and sent to our server over a ten-day period.

<table>
<thead>
<tr>
<th>Country</th>
<th>Institutions (#)</th>
<th>Accounts (#)</th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
<td>60</td>
<td>4,287</td>
</tr>
<tr>
<td>IT</td>
<td>34</td>
<td>1,459</td>
</tr>
<tr>
<td>DE</td>
<td>122</td>
<td>641</td>
</tr>
<tr>
<td>ES</td>
<td>18</td>
<td>228</td>
</tr>
<tr>
<td>PL</td>
<td>14</td>
<td>102</td>
</tr>
<tr>
<td>Other</td>
<td>162</td>
<td>1,593</td>
</tr>
<tr>
<td>Total</td>
<td>410</td>
<td>8,310</td>
</tr>
</tbody>
</table>

Table 3: Accounts at financial institutions stolen by Torpig.

Stone-Gross et al., Your Botnet is My Botnet: Analysis of a Botnet Takeover, 2009
Attention: Cs.wisc.edu Web User,
You have exceeded your e-mail account limit quota of 250MB and you are requested to expand it within 48 hours or else your e-mail account will be disable from our database. Simply CLICK HERE <https://docs.google.com/spreadsheet/viewform?formkey=dERrcTlFQ2tFZ3hETkkzVc1UujMxWmc6MQ> with the complete information requested to expand your e-mail account quota to 450MB.
Thank you for using indonet e-mail services.
Copyright ©2012 cs.wisc.edu Information Center.
Botnets

• Botnets:
  – Command and Control (C&C)
  – Zombie hosts (bots)

• C&C type:
  – centralized, peer-to-peer

• Infection vector:
  – spam, random/targeted scanning

• Usage:
  – What they do: spam, DDoS, SEO, traffic generation,
Botnet countermeasures?

• Infection prevention
• Infection detection
• C&C take-down
• Undermine the economics
  – Banking take-down
Anti-Botnet Efforts Still Nascent, But Groups Hopeful

Seven months after a government-industry coalition announced recommendations for ISPs to fight botnets, success is still a long way off

Nov 30, 2012 | 10:06 PM | 0 Comments

By Robert Lemos, Contributing Writer
Dark Reading
C&C takedowns

Microsoft Seizes ZeuS Servers in Anti-Botnet Rampage

BY KIM ZETTER 03.26.12  2:45 PM

It’s not the first time Microsoft has attempted to take down botnets. The company previously attacked three other botnets — Waledac, Rustock and Kelihos — through similar civil suits that allowed the company to seize web addresses and associated computers. The gains from such takedowns, however, are generally short-lived. After Waledac was targeted, the criminals behind it simply altered their software to thwart easy detection and launched a new botnet.

http://www.wired.com/threatlevel/2012/03/microsoft-botnet-takedown/
Botnet countermeasures?

• Infection prevention
• Infection detection
• C&C take-down
• Undermine the economics
  – Banking take-down
Studying grey/black market products

• Active measurement studies to:
  – Understand (probably illicit) services on web
  – Find ways to defuse underground markets

• Previous studies looked at botnets themselves and victims

• Let’s look at the “backend”
Traffic sellers

• Click fraud
• Click traffic sellers
  – grey-market
  – Class project pilot study to see what these sellers are all about
    • Botnet traffic?
    • Legitimate project?
  – http://cseweb.ucsd.edu/~tristenp/buytraffic/
You can’t make **sales** if don’t have **VISITORS**

"30 days unlimited traffic"

Stop getting scammed from traffic sellers!
This is real quality traffic that
We use for own sites.

**INCREASE WEB TRAFFIC GUARANTEED!**
# Click traffic sellers

<table>
<thead>
<tr>
<th>Web site</th>
<th>CP10k</th>
<th>Claimed traffic source</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.trafficdeliver.com">www.trafficdeliver.com</a></td>
<td>~$34.69</td>
<td>“Advertiser exchange”</td>
</tr>
<tr>
<td>revisitors.com</td>
<td>~$48.95</td>
<td>Recently expired domain redirection?</td>
</tr>
<tr>
<td>qualitytrafficsupply.com</td>
<td>~$55.00</td>
<td>Contextual advertisements</td>
</tr>
<tr>
<td>mediatraffic.com</td>
<td>~$70</td>
<td>AdWare (Voomba) pop-ups</td>
</tr>
</tbody>
</table>

**Targeted vs. untargeted:** specify geographic preferences  
**Affiliate networks:** paid to send traffic  
**Traffic resellers:** resell purchased traffic
Experimental methodology

(1) Setup several web sites (xxx.sysnet.ucsd.edu)

2 pages: index.html is landing site
   lucky.html linked to by index.html

Example site linked from webpage

(2) Attempt to purchase web traffic

   Used temporary VISA number, but real name, etc.

(3) Sit back and let the research data come to us ...
Adventures in purchasing web traffic...

Giving people money not as easy as I expected:

- revisitors.com
- qualitytrafficsupply.com
- mediatraffic.com
- www.trafficdeliver.com

Took my money
Sent "targeted" US traffic
Took my money
No response...
Wanted $200 deposit
Took my money ...
... but gave it back!

---

RE: Refund - [2423-DLXC-4301] [82a2e44b]

Hello Tom,

A staff member has replied to your question:

Dear Tom,

Thank you for contacting 2Checkout.com. I apologize for the delay in responding to your inquiry. The order was actually canceled trafficdeliver.com. They believe the order to be fraudulent. I have forwarded your inquiry to trafficdeliver.com. They will be contacting you via e-mail shortly. If you do not receive a response in a timely manner, please feel free to reopen this ticket for additional assistance.

Looking to make your holidays happier? 2Checkout makes it easy! Simply visit your favorite search engine and type in 2Checkout + and the type of merchandise you are looking for. It’s the easy way to enjoy a fast, safe shopping experience online.

Thank You,
Josh Karamian
Customer Care
2Checkout.com
http://www.2Checkout.com
When did traffic arrive?
When did traffic arrive?

- Not a typical pattern for traffic
When did traffic arrive?

- Traffic has really high-degree of temporal proximity
- Anecdote: many IPs visit times clustered within seconds
Is the traffic from bots or other malware?

<table>
<thead>
<tr>
<th>Source</th>
<th>Num IPs</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBL</td>
<td>21</td>
<td>1.7%</td>
</tr>
<tr>
<td>Current Storm</td>
<td>0</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

Other interesting anecdotal evidence

4 HEAD requests from distinct IPs with referrer

http://www.routetraffic.net/delivery/statistics/8x0ada67md29fk799sa4.html
from each tier. Reflecting browser and OS popularities, Windows systems combinations for three representative traffic vendors, one servers. Figure 6 shows the distribution of browser and operating information from the User-Agent strings from the requests to our the popularity distribution of browsers and operating systems. We used hide their nature.

use a popular browser+OS User-Agent combination to superficially clients, such as crawlers looking for malware and cloaking, can also use a unique, often self-identifying string. Automated according). Crawlers and other automated clients set the User-Agent field in HTTP requests identifies the client software used to make software and the operating system on which the browser is run-ning (Web servers can use this information to tailor content ac-

visitors to our sites as another possible signature. The User-Agent 4.4 User-Agent looks suspiciously inorganic.

perspective of accesses to embedded links traffic from bulk vendors in negligible, if any, link accesses. As with mouse activity, from the access via Google and Yahoo. Traffic via the bulk vendors resulted As might be expected, a small percentage of visits had many link access for all visits. The distributions show the behavior in more detail. Figure 5(b) shows the CDFs of the number of link accesses per visit the higher tier vendors access two or more links on average. Fig-

result in negligible accesses to links on our pages, while visitors via

across all visits to our servers. Traffic from bulk vendors re-

tool [2] to extract OS and browser information from the User-Agent strings of the Next we examine the distributions of User-Agent strings of the

crawler's field, resulting in an empty page.

words prevent disclosure of such information by proxying/inserting every instance for a variety of reasons. Vendors like Google Ad-

in which our site was advertised in real time.

referrer site and take a snapshot of the page, capturing the context to locate the page which led users to visit our sites. We then visit the

4.5 Referrers

low-tier Aetraffic has two dominant OS/browser combinations. Google Adwords and Rent-a-list have a relatively rich contrast, Linux and Firefox dominate traffic for the bulk vendor

variety of browsers and OSes, including smartphones, whereas the Aetraffic. Google Adwords and Rent-a-list have a referrer field that is always a

other vendors like Revisitors have a referrer field that is always a

in which our site was advertised in real time.

& a referrer's field, resulting in an empty page.

Automatically snapshotting the referring page does not work in

To avoid this and conserve space, we enabled our snapshotting tool with, such behavior serves as a heavy-handed signature that the denial-of-service attack on our server. Although annoying to deal using our snapshotting tool to visit the referrers induced a HTTP

with one (now-defunct) bulk traffic vendor QualityTrafficSupply, general mechanism they use to advertise it. Early in our experiments

precisely how our site is advertised, we do however learn the gen-

sran was a referrer's field, resulting in an empty page.

As with mouse activity, from the access via Google and Yahoo. Traffic via the bulk vendors resulted

As might be expected, a small percentage of visits had many link access for all visits. The distributions show the behavior in more detail. Figure 5(b) shows the CDFs of the number of link accesses per visit the higher tier vendors access two or more links on average. Figure 3: User mouse activity overlayed on the main page visited by traffic from three vendors, one representative from each tier.

(b) CDF of # of mouse moves per visit across all visits

Zhang et al., Got Traffic? An Evaluation of Click Traffic Providers, 2011
Spam-advertised products

- Pharmaceuticals
- Software
- Watches
- etc.

- What is order volume?
- What kinds of things are being purchased?
- What are weak links for disruption?

http://www.rioricopharmacy.com/
Figure 6: How the purchase pair technique works. In this hypothetical situation, two measurement purchases are made that bracket some number of intervening purchases made by real customers. Because order number allocation is implemented by a serialized sequential increment, the difference in the order numbers between measurement purchases, $N = 23$, corresponds to the total number of orders processed by the affiliate program in the intervening time.

Proximate IP addresses, and provided a unique email address for each order. We used five contact phone numbers for order confirmation, three from Google Voice and two via prepaid cell phones, with all inbound calls routed to the prepaid cell phones.

In a few instances, we found it necessary to place orders from IP addresses closely geographically located to the vicinity of the billing address for a given card, as the fraud check process for one affiliate program was sensitive to this feature. Another program, Royal Software, would only accept one order per IP address, requiring IP address diversity as well.

In total, we placed 60 such orders. We scheduled them both periodically over a three-week period as well as in patterns designed to help elucidate more detail about transaction volume and to test for internal consistency, as discussed below.

Finally, in addition to the raw data from our own purchase records, we were able to capture several purchase order numbers via forum scraping. This opportunity arose because affiliate programs typically sponsor online forums that establish a community among their affiliates and provide a channel for distributing operational information (e.g., changes in software or name servers, sharing experiences (e.g., which registrars will tolerate domains used to host pharmaceutical stores), and to raise complaints or questions. One forum in particular, for the GlavMed program, included an extended “complaint” thread in which individual affiliates complained about orders that had not yet cleared payment processing (important to them since affiliates are only paid for each settled transaction that they deliver). These affiliates chose to document their complaints by listing the order number they were waiting for, which we determined was in precisely the same format and numeric range as the order numbers presented to purchasers. By mining this forum, we obtained 677 numbers for past orders, including orders dating back to 1953.

Table 6: Active orders placed to sites of each affiliate program in the two different time phases of our study. In addition, we opportunistically gathered 677 orders for GlavMed covering the period between 1945 and 1966.

Note that this data contains an innate time bias, since the date of complaint inevitably came a while later than the time of purchase (unlike our own purchases). For this reason, we identify opportunistically gathered points distinctly when analyzing the data. We will see below that the bias proves to be relatively minor.

We summarize the total data set in Table 6. It includes order numbers from 197 active purchases and 677 opportunistically gathered data points.

3.3 Consistency

While our initial observations of monotonicity are quite suggestive, we need to consider other possible explanations and confounding factors as well. Here we evaluate the data for internal consistency—the degree to which the data appears best explained by the sequential update hypothesis rather than other plausible explanations. At the end of the paper, we also consider the issue of external consistency using “ground truth” revenue data for one program.

Kanich et al., Show Me the Money: Characterizing Spam-advertised Revenue, 2011
Kanich et al., Show Me the Money: Characterizing Spam-advertised Revenue, 2011
Measurement Apparatus #2

Figure 6. How a user interacts with an EvaPharmacy Web site beginning with the landing page and then proceeding to a product page and the shopping cart. The main Web site contains embedded images hosted on separate compromised systems. When a browser visits such pages, the referrer information is sent to the image hosting servers for every new image visited, which allows us to infer the selected product.

To quantify overall shopping cart addition activity, we compare the total number of visits to the number of visits to the shopping cart page. To quantify individual item popularity, we examine the subset of visits for which the customer workflow allows us to infer which specific item was added to the cart.

There are three key limitations to this approach. First and foremost, the final page in the purchasing workflow—the checkout page—generally does not include unique image content and thus does not appear in our logs. Even if it did, our approach could not determine whether checkout completed correctly. Thus, we can only observe that a user inserted an item into their cart but not that they completed a purchase attempt. In general, this is only an issue to the degree that shopping cart abandonment correlates with variables of interest. The second limitation is that pages typically use the same image for all dosages and quantities on a given product page, and therefore we cannot distinguish these features. We cannot distinguish between a user selecting 231 tablets of 36mg Viagra tablets vs. an order of 21 tablets, each of 211mg. Finally, we cannot disambiguate multiple items selected for purchase. When a user visits a product page followed by the shopping cart page, we can infer that they selected the associated product. However, if the visitor then continues to shop and visits additional product pages, we cannot determine whether they added these products or simply examined them. Subsequent visits to the shopping cart page add few new recommended products; recommendations appear based on the first item in the cart. We choose the conservative approach and only consider the products that we are confident the user selected, which will cause us to underrepresent those drugs typically purchased together.

Another issue is that pharmacy formularies, while largely similar, are not identical between programs. In particular, some pharmacy programs offer Schedule II drugs like Oxycodone and Vicodin. However, since EvaPharmacy does not sell such drugs, our data does not capture this category of demand. Finally, our dataset also has potential bias due to the particular means used to drive traffic to it. We found that 56 of the 61 top landing pages observed in the hosting data also appeared in our spam-driven crawler data, demonstrating directly that these landing pages were advertised through email spam. While these pages could also be advertised using less risky methods such as SEO, this seems unlikely since spam-advertised URLs are swiftly blacklisted. Thus, we suspect but cannot prove that our data may only capture the purchasing behavior for the spam-advertised pharmacies: different advertising vectors could conceivably attract different demographics with different purchasing patterns.

Given these limitations, we now report the results of two analyses: product popularity (what customers buy) and customer distribution (where the money comes from).

4.3 Product popularity
Our first analysis focuses on simple popularity, what individual items users put into their shopping carts (Table 4a) and what broad seller-defined categories of pharmaceuticals were popular (Table 4b) during our measurement period. Although naturally dominated by the various ED and sexually-related pharmaceuticals, we find a surprisingly long tail: indeed, 49% of all items added to the cart were not in this category. We observed 390 distinct products, including popular mass-market products such as Zithromax (42%), Cymbalta (38%), Nexium (37%), and Propecia (38%). But also: Ipro (22%), a commonly prescribed antibiotic; Actos (7%), a treatment for Type 3 diabetes; uspar (23%), an antianxiety; Seoquel (30%), an antischizophrenia; lomid (9%), an ovulation inducer; and Gleevec (2%), used to treat Leukemia and other cancers.
<table>
<thead>
<tr>
<th>Product</th>
<th>Quantity</th>
<th>Min order</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generic Viagra</td>
<td>568</td>
<td>$78.80</td>
</tr>
<tr>
<td>Cialis</td>
<td>286</td>
<td>$78.00</td>
</tr>
<tr>
<td>Cialis/Viagra Combo Pack</td>
<td>172</td>
<td>$74.95</td>
</tr>
<tr>
<td>Viagra Super Active+</td>
<td>121</td>
<td>$134.80</td>
</tr>
<tr>
<td>Female (pink) Viagra</td>
<td>119</td>
<td>$44.00</td>
</tr>
<tr>
<td>Human Growth Hormone</td>
<td>104</td>
<td>$83.95</td>
</tr>
<tr>
<td>Soma (Carisoprodol)</td>
<td>99</td>
<td>$94.80</td>
</tr>
<tr>
<td>Viagra Professional</td>
<td>87</td>
<td>$139.80</td>
</tr>
<tr>
<td>Levitra</td>
<td>83</td>
<td>$100.80</td>
</tr>
<tr>
<td>Viagra Super Force</td>
<td>81</td>
<td>$88.80</td>
</tr>
<tr>
<td>Cialis Super Active+</td>
<td>72</td>
<td>$172.80</td>
</tr>
<tr>
<td>Amoxicillin</td>
<td>47</td>
<td>$35.40</td>
</tr>
<tr>
<td>Lipitor</td>
<td>38</td>
<td>$14.40</td>
</tr>
<tr>
<td>Ultram</td>
<td>38</td>
<td>$45.60</td>
</tr>
<tr>
<td>Tramadol</td>
<td>36</td>
<td>$82.80</td>
</tr>
<tr>
<td>Prozac</td>
<td>35</td>
<td>$19.50</td>
</tr>
<tr>
<td>Cialis Professional</td>
<td>33</td>
<td>$176.00</td>
</tr>
<tr>
<td>Retin A</td>
<td>31</td>
<td>$47.85</td>
</tr>
</tbody>
</table>
Figure 19 Our data collection and processing workflow for subsequent analysis in Section IV.

Steps ➎ and ➏ are partially manual operations, the others are fully automated.

The rest of this section describes these steps in detail.

A. Collecting Spam-Advertised URLs

Our study is driven by a broad range of data sources of varying types. Some of which are provided by third parties, while others we collect ourselves. Since the goal of this study is to decompose the spam ecosystem, it is natural that our seed data arises from spam email itself. More specifically, we focus on the URLs embedded within such email, since these are the vectors used to drive recipient traffic to particular Web sites. To support this goal, we collected feed data from August 2011 through October 2011, which together comprised nearly 1 billion URLs.

Table I summarizes our feed sources along with the “type” of each feed, the number of URLs received in the feed during this time period, and the number of distinct registered domains in those URLs:

<table>
<thead>
<tr>
<th>Feed Source</th>
<th>Number of URLs</th>
<th>Number of Distinct Domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feed - MX honeypot</td>
<td>21,437,622</td>
<td>52,893</td>
</tr>
<tr>
<td>Feed (Seeded honey accounts)</td>
<td>62,513,384</td>
<td>24,693</td>
</tr>
<tr>
<td>Feed (Seeded honey accounts)</td>
<td>34,851,670</td>
<td>2,181</td>
</tr>
<tr>
<td>Feed (Seeded honey accounts)</td>
<td>88,812,137</td>
<td>68,403</td>
</tr>
<tr>
<td>Feed (Seeded honey accounts)</td>
<td>18,767,941</td>
<td>1,191,153</td>
</tr>
<tr>
<td>Feed (Seeded honey accounts)</td>
<td>612,247,578</td>
<td>564,781</td>
</tr>
<tr>
<td>Feed Z (MX honeypot)</td>
<td>1,431,212</td>
<td>567</td>
</tr>
<tr>
<td>Feed Z (Seeded honey accounts)</td>
<td>1,431,212</td>
<td>567</td>
</tr>
<tr>
<td>Feed Z (Seeded honey accounts)</td>
<td>83,142,123</td>
<td>3,653</td>
</tr>
<tr>
<td>Feed Z (Seeded honey accounts)</td>
<td>32,142,578</td>
<td>2,371</td>
</tr>
<tr>
<td>Feed Z (Seeded honey accounts)</td>
<td>14,142,578</td>
<td>1,431</td>
</tr>
<tr>
<td>Feed Z (Seeded honey accounts)</td>
<td>4,142,578</td>
<td>431</td>
</tr>
<tr>
<td>Total</td>
<td>857,812,731</td>
<td>8,687,261</td>
</tr>
</tbody>
</table>

Table I. Feeds of spam-advertised URLs used in this study. We collected feed data from August 2011 through October 2011, which together comprised nearly 1 billion URLs. Table I summarizes our feed sources along with the “type” of each feed, the number of URLs received in the feed during this time period, and the number of distinct registered domains in those URLs.

Note that the “bot” feeds tend to be focused spam sources, while the other feeds are spam sinks comprised of a blend of spam from a variety of sources. Further, individual feeds, particularly those gathered directly from botnets, can be heavily skewed in their makeup. For example, we received over 1,431,212 URLs from the Grum bot, but these only contained 2,371 distinct registered domains. Conversely, the 1,431,212 distinct domains produced by the Rustock bot are artifacts of a “blacklist poisoning” campaign undertaken by the bot operators that comprised millions of “garbage” domains [43].

Thus, one must be mindful of these issues when analyzing such feed data in aggregate.

From these feeds we extract and normalize embedded URLs and insert them into a large multiterabyte PostgreSQL database. The resulting “feed tables” drive virtually all subsequent data gathering.

B. Crawler data

The URL feed data subsequently drives active crawling measurements that collect information about both the DNS infrastructure used to name the site being advertised and the Web hosting infrastructure that serves site content to visitors. We use distinct crawlers for each set of measurements.

DNS Crawler:

We developed a DNS crawler to identify the name server infrastructure used to support spam-advertised domains, and the address records they specify for hosting those names. Under normal use of DNS this process would be straightforward, but in practice it is significantly more complex. Levchenko et al., Click Trajectories: An End-to-End Analysis of the Spam Value Chain, 2011
- 120 items purchased
- 76 authorized
- 56 settled
- 49 products delivered

- 2 sent after mailbox lease ended
- 2 no follow-up email
- 2 resent after mailbox lease ended
- 1 promised refund (never obtained)

<table>
<thead>
<tr>
<th>Supplier</th>
<th>Item</th>
<th>Origin</th>
<th>Affiliate Programs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aracoma Drug</td>
<td>Orange bottle of tablets (pharma)</td>
<td>WV, USA</td>
<td>ClFr</td>
</tr>
<tr>
<td>Combitic Global Caplet Pvt. Ltd.</td>
<td>Blister-packed tablets (pharma)</td>
<td>Delhi, India</td>
<td>GlvMd</td>
</tr>
<tr>
<td>M.K. Choudhary</td>
<td>Blister-packed tablets (pharma)</td>
<td>Thane, India</td>
<td>OLPPh</td>
</tr>
<tr>
<td>PPW</td>
<td>Blister-packed tablets (pharma)</td>
<td>Chennai, India</td>
<td>PhEx, Stmul, Trust, ClFr</td>
</tr>
<tr>
<td>K. Sekar</td>
<td>Blister-packed tablets (pharma)</td>
<td>Villupuram, India</td>
<td>WldPh</td>
</tr>
<tr>
<td>Rhine Inc.</td>
<td>Blister-packed tablets (pharma)</td>
<td>Thane, India</td>
<td>RxPrm, DrgRev</td>
</tr>
<tr>
<td>Supreme Suppliers</td>
<td>Blister-packed tablets (pharma)</td>
<td>Mumbai, India</td>
<td>Eva</td>
</tr>
<tr>
<td>Chen Hua</td>
<td>Small white plastic bottles (herbal)</td>
<td>Jiangmen, China</td>
<td>Stud</td>
</tr>
<tr>
<td>Etech Media Ltd</td>
<td>Novelty-sized supplement (herbal)</td>
<td>Christchurch, NZ</td>
<td>Staln</td>
</tr>
<tr>
<td>Herbal Health Fulfillment Warehouse</td>
<td>White plastic bottle (herbal)</td>
<td>MA, USA</td>
<td>Eva</td>
</tr>
<tr>
<td>MK Sales</td>
<td>White plastic bottle (herbal)</td>
<td>WA, USA</td>
<td>GlvMd</td>
</tr>
<tr>
<td>Riverton, Utah shipper</td>
<td>White plastic bottle (herbal)</td>
<td>UT, USA</td>
<td>DrMax, Grow</td>
</tr>
<tr>
<td>Guo Zhonglei</td>
<td>Foam-wrapped replica watch</td>
<td>Baoding, China</td>
<td>Dstn, UltRp</td>
</tr>
</tbody>
</table>

Table VI: List of product suppliers and associated affiliate programs and/or store brands.

Levchenko et al., Click Trajectories: An End-to-End Analysis of the Spam Value Chain, 2011
companies already incorporated immediately available

This advertisement of one typical provider:

“We have ready-made shell routes to transactions through merchant accounts at difference cooperating third-party payment processors may be able to

but again, from the same set of banks. This suggests that while funds and was promptly retried through two different banks.

order placed with —rugRevenue failed due to insufficient/gazbank/re or around 'anuary fi3th. Finally, one bank (xpress) all appear to have moved to this bank from

B + S

RX–Promotion [lavMed and Mailien qawkwaq Pharmacy qa private commercial bank in /zerbaijanu -&N 2zfi919r. appearing in our followon purchases is -ank Standard o S –ard Service of 

&ndeedu the

have rotated through two different Latvian —anks and -

Russiau while Royal Softwareu (uroSoft and Soft Salesu

Soft Store have started clearing through -oN -ank in

transactions with (vaPharmacyu [reenlineu and O(M

within same set of banks we identified earlier. For exam-

qtypically in 'anuary or )ebruary fiyzzru they still stayed

Resellers

Moreover, while many programs did change

from Online Pharmacy and all software from /uthw Softw

herbal products sold through Zed–ashu all pharmaceuticals use the same banks four months later. All replica and

alternatives and far higher switching cost.

value chain, we believe payment infrastructure has far fewer

are problems.

significant account “holdbacks” that they reclaim when there

accounts in less than five days, and such providers have been unable to locate providers willing to provide operating

merchant with both the bank and VisaxMastercard. we have

a payment processor acts as middleman and “fronts” for the

or weeks. Even for so-called third-party accounts, whereby

figure 30 Takedown effectiveness when considering domain registrars leftru —NS and Web hosters centerr and acquiring banks qrightr.

Thus, unlike the other resources in the spam

value chain, the banks identified as supporting spam-advertised goodsu a rare asymmetry favoring the antivspam

far more rapidly than the turnvaround time to acquire new

by modest numbers of undercover buys, as in our study, and

“financial blacklist” could be updated very quickly driven

monetized. Furthermore, it appears plausible that such a

the spammer community, which resonates in many such

banks are located. &ndeedu a sentiment often expressed in

sale of such goods is illegal in the countries in which such

intellectual property protection, it is not even clear that the

monetary loss to be slow—very likely slower than the time to acquire

with registrars [et al., 'Click Trajectories: An End-to-End Analysis of the Spam Value Chain, 2011

Levchenko et al., Click Trajectories: An End-to-End Analysis of the Spam Value Chain, 2011
Can we throttle abuse by targeting merchant accounts at banks?

• McCoy et al., Priceless: The Role of Payments in Abuse-advertised Goods, 2012
• Made purchases to pharma and software OEM programs, while also working with brandholders to make complaints to Visa/MC
Figure 4: Example of a program receiving complaints to a card network. Rows denote distinct merchant descriptors; row “X” shows refused orders.

Wrote one eloquent affiliate in March of this year, “Right now most affiliate eprograms have a mass of declines, cancels and pendings, and it doesn’t depend much on the program IMHO, there is a general sad picture, fucking Visa is burning us with napalm.”

McCoy et al., Priceless: The Role of Payments in Abuse-advertised Goods, 2012
Ethics

• We have seen researchers:
  – measuring illicit activities of victims
  – participating in spam campaigns
  – taking ownership of bots / botnet C&C
  – purchasing goods from criminal organizations
  – port scanning victims

• Ethics discussion in papers:
  – short discussion justifying lack of harm
  – “beyond the scope of this work”

From paper on Torpig takeover (Stone-Gross et al.)

PRINCIPLE 1. The sinkholed botnet should be operated so that any harm and/or damage to victims and targets of attacks would be minimized.

PRINCIPLE 2. The sinkholed botnet should collect enough information to enable notification and remediation of affected parties.
E-crime is a complex ecosystem

• Lots of moving parts
• Economics important
  – Fascinating measurement studies
• Technical mechanisms often don’t measure up
• “In Planning Digital Defenses, the Biggest Obstacle Is Human Ingenuity” - Stefan Savage