The OTHER spam problem:
spam as evidence, the UAB Spam Data Mine,
and UAB PhishIntel™

January 21, 2011
Gary Warner
UAB
Objective

- UAB’s Computer Forensics Research Laboratory
- The “Other” Spam Problem
- The UAB Spam Data Mine example:
  - Building spam clusters for evidence
- PhishIntel Demo
- Future problems
The Mission

• UAB Computer Forensics

• A partnership between Computer Science and Justice Science

• Using our team of Computer Scientists and Criminologists to Research & Develop Tools, Training, and Techniques to improve our response to Cybercrime

• Producing the most qualified graduates for cybercrime investigator positions

• Raising Awareness through public education of current cybercrime trends and counter-measures
Research Areas

• The UAB Spam Data Mine
• The UAB PhishURLs Project
• Malware Investigations
• Cyber Crime Investigations
• Cyber Intelligence

• http://www.cis.uab.edu/forensics/
Everyone’s Spam Problems

• Typical User: I get spam and I don’t want to.
• Typical SysAdmin: The spam filter just ate my boss’s report
• Typical Researcher: My Hidden Markov Model naïve Bayesian filter currently filters 99.115% of spam and with my new tweaks it will filter 99.13% while maintaining a low false positive rate
True confession

I don’t care about any of those.
My Spam Problem

• In the United States, it is illegal:
  – to send email with false or misleading headers
  – to send email advertising controlled substances
  – to send email as part of a conspiracy to defraud
  – to plant malware on someone’s computer to make them send email
  – to send email as part of a conspiracy to infect computers with malware
  — BUT NOBODY IS GOING TO JAIL!!!
My Spam Problem

• The entire focus of the UAB Spam Data Mine is to provide evidence to law enforcement and other investigators that will allow them to document, investigate, identify, indict, prosecute, and convict violations and violators of cybercrime law.

• If, as a side effect, others might benefit from that data to assist with THEIR spam problems, we are happy to collaborate.
Perspective #1

Gar: What do you do with your filtered spam?
X: We throw it away
Gar: In my line of work, we call that “destruction of evidence”
Perspective #2

X: We don’t have a spam problem.
Gar: So . . . All the people who spammed you are now in jail?
Cybercrime and Society

- Cybercrime is not a TECHNOLOGY problem.
- Cybercrime is a SOCIETAL problem.

- Why did you decide to pursue your current career instead of being a professional bank robber?

- There must be clear messaging that when you commit a cybercrime, you will be caught, and you will pay an appropriate penalty.
Victims Rights

• A Victims Rights advocate who specialized in identity theft victimization, working closely with the Office of Justice Programs.
• He asked “What do Identity Theft Victims Want Most?”
• He says I’m one of the few that got the answer right:

JUSTICE
Will your Spam Case Be Prosecuted?

Here’s a simple questionnaire to determine whether your spam case will be prosecuted:

Q1: Are you currently a candidate for the office of Vice President of the United States?
   – Yes: Proceed to Question 2.
   – No: Thanks for your call. Good-bye.
From Dec 24 to January 8, Rustock stopped spamming.

One botnet can make that much difference to spam levels.

During that time his botnet did clickfraud instead. 8-)}
I’m at Google, so I should totally skip this slide. 8-(

In law enforcement circles when I mention that we have 144 pentium cores and 90 TB of storage dedicated to cybercrime datamining, people are impressed.

I know that doesn’t work here. Moving on. 8-(
Spam Collection vs Spam Data Mine

• Misconception:
  – To search the UAB Spam Data Mine, LE gives us a keyword and give them back emails containing that keyword.

• Reality:
  – Give us a sample EMAIL, and we’ll show you various forms of “relatedness” to other emails.
Manual Spam Clustering

• An explanation of how we would “manually” use the UAB Spam Data Mine for clustering can be found in our technical report:
  • UABCIS-TR-2010-120510 - Manual Clustering in the UAB Spam Data Mine (http://www.cis.uab.edu/forensics/TechReports)

• We’ll walk through the process on the next few slides
“Sender” vs. “Target” clusters

- We know that there are several levels of infrastructure to spam. We’ll take the example of a spam affiliate program.
- On the “Sender” side, we have several options:
  - Lease rack space to send my spam
  - Create a botnet to send my spam
  - Rent “time” on another botnet to send my spam
UAB Spam Clustering Papers


“UGG” Rackspace spammers
(Dec 2010)

• 173.208.82.0/24 has 13 different IP addresses which sent me between 20 and 315 messages each
• each IP spamming a different destination address, and each using a consistent but unique sender domain.
• Each IP used ONE from, ONE destination, and ONE subject, but each was different.
• That netblock is Ubiquity Servers of Chicago subleasing to Nobis Technology Group of Phoenix, subleasing to Server Results of Carson City, Nevada.
“UGG” Rackspace spammers (Dec 2010)

- Twelve IPs between 174.127.69.98-.184
  - Providence Hosting - midphase.com - Providence, Utah

- Nine IPs between 206.71.51.21-.44
  - Galaxy Visions - rightcircular.com - Brooklyn, NY

- Thirteen IPs between 206.71.57.33-.59
  - Galaxy Visions - urscredit.com - Brooklyn, NY

- Eight IPs between 207.244.219.195-.222 - US Net Inc
  - hostirian.com - St. Louis, MO

- Ten IPs between 209.200.50.69-.122
  - WebAir - sandplus.com - Westbury, NY

- Eight IPs between 209.250.225.83-.94
  - Seacaucus Rackvibe - Secaucus, NJ

- Eleven IPs between 67.18.183.98-.126
  - The Planet - theplanet.com - Dallas, TX

- Eight IPs between 68.67.85.113-.253
  - 3dgwebhosting.com - momslovetosave.com - Las Vegas, NV
DEA: a “Botnet” spammer example (from December 5, 2010)
Other subjects with same destination domain name

```
iiidspam=# select distinct subject from decspam where message_id in (select message_id from declink where machine = 'cheaprx-ra.com');
```

<table>
<thead>
<tr>
<th>Table &quot;public.decsapm&quot;</th>
<th></th>
<th>Type</th>
<th>Table &quot;public.declink&quot;</th>
<th></th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Column</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>message_id</td>
<td>text</td>
<td></td>
<td>message_id</td>
<td>text</td>
<td></td>
</tr>
<tr>
<td>subject</td>
<td>text</td>
<td></td>
<td>machine</td>
<td>text</td>
<td></td>
</tr>
<tr>
<td>sender_name</td>
<td>text</td>
<td></td>
<td>path</td>
<td>text</td>
<td></td>
</tr>
<tr>
<td>sender_username</td>
<td>text</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sender_domain</td>
<td>text</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sender_ip</td>
<td>cidr</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>receiving_date</td>
<td>date</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>time_stamp</td>
<td>time with time zone</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>time_in_text</td>
<td>text</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>text_length</td>
<td>integer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>word_count</td>
<td>integer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>subject_md5</td>
<td>text</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sender_name_md5</td>
<td>text</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(35 rows)
“Simple matching”

```sql
select count(distinct message_id) from decspam where message_id in (select message_id from declink where machine = 'cheaprx-ra.com');

count
------
 302
(1 row)
```

302 email messages in our December 1st-5th spam advertised the domain “cheaprx-ra.com”

```sql
select count(distinct sender_ip) from decspam where message_id in (select message_id from declink where machine = 'cheaprx-ra.com');

count
------
 270
(1 row)
```

270 computers in our December 1st-5th spam sent those emails.
Same subject – other destinations

• 9,562 spam emails in December used the same subjects as the emails for “cheaprx-ra.com”
• Those subjects advertised 985 additional hostnames on 68 unique domains other than “cheaprx-ra.com”
• (Repeating the experiment for the full month of November, there were 213 domains identified)
### Same Sender-IP Clusters

A nested query can give us a list of all the OTHER spam sent from the same IP addresses that sent us the ‘cheaprx-ra.com’ spam.

<table>
<thead>
<tr>
<th>IP Address</th>
<th>Message Description</th>
<th>Domain Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>24.244.180.6/32</td>
<td>Customer profile confirmation</td>
<td>wish3.exclusiverefill.ru</td>
</tr>
<tr>
<td>41.140.126.247/32</td>
<td>Customer shipping confirmation</td>
<td>cheaprx-ra.com</td>
</tr>
<tr>
<td>46.187.22.21/32</td>
<td>Order status</td>
<td>cheaprx-ra.com</td>
</tr>
<tr>
<td>41.140.126.247/32</td>
<td>Customer shipping update</td>
<td>cheaprx-ra.com</td>
</tr>
<tr>
<td>46.187.22.21/32</td>
<td>Customer shipping confirmation</td>
<td>cheaprx-ra.com</td>
</tr>
<tr>
<td>46.187.22.21/32</td>
<td>Reorder re minder confirmation</td>
<td>trusted-rxsi.com</td>
</tr>
<tr>
<td>46.187.22.21/32</td>
<td>Reorder reminder</td>
<td>trusted-rxsi.com</td>
</tr>
<tr>
<td>62.105.25.176/32</td>
<td>Customer update on order</td>
<td>cheaprx-ra.com</td>
</tr>
<tr>
<td>62.105.25.176/32</td>
<td>Special discount update</td>
<td>totrusted-rx.com</td>
</tr>
<tr>
<td>62.105.25.176/32</td>
<td>Customer shipping confirmation</td>
<td>trusted-rxol.com</td>
</tr>
<tr>
<td>62.105.25.176/32</td>
<td>Urgent, re fill reminder</td>
<td>cheaprx-ra.com</td>
</tr>
<tr>
<td>62.105.25.176/32</td>
<td>Order status reminder</td>
<td>totrusted-rx.com</td>
</tr>
<tr>
<td>62.105.25.176/32</td>
<td>Reorder reminder</td>
<td>trgt.pillsfillrefill.ru</td>
</tr>
<tr>
<td>62.105.25.176/32</td>
<td>Customer profile update</td>
<td>trusted-rxal.com</td>
</tr>
<tr>
<td>62.105.25.176/32</td>
<td>Customer update, urgent</td>
<td>trusted-rxsi.com</td>
</tr>
<tr>
<td>62.105.25.176/32</td>
<td>Customer profile confirmation</td>
<td>vigra.genuinemy.ru</td>
</tr>
<tr>
<td>62.105.25.176/32</td>
<td>Reorder reminder</td>
<td>totrusted-rx.com</td>
</tr>
<tr>
<td>67.215.25.185/32</td>
<td>Special discount confirmation</td>
<td>cheaprx-ra.com</td>
</tr>
</tbody>
</table>

• (Technical report contains full listing)
Same Destination Clusters

- select a.subject, b.machine from decspam a, declink b where a.message_id = b.message_id and a.subject = 'Order status';

<table>
<thead>
<tr>
<th>subject</th>
<th>machine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Order status</td>
<td>trusted-rxal.com</td>
</tr>
<tr>
<td>Order status</td>
<td>trusted-rxal.com</td>
</tr>
<tr>
<td>Order status</td>
<td>trusted-rxsi.com</td>
</tr>
<tr>
<td>Order status</td>
<td>sss.toprefilli.ru</td>
</tr>
<tr>
<td>Order status</td>
<td>sss.toprefilli.ru</td>
</tr>
<tr>
<td>Order status</td>
<td>trusted-rxol.com</td>
</tr>
<tr>
<td>Order status</td>
<td>wotrusted-rx.com</td>
</tr>
<tr>
<td>Order status</td>
<td>tottrusted-rx.com</td>
</tr>
<tr>
<td>Order status</td>
<td>ssa.pillsagent.ru</td>
</tr>
<tr>
<td>Order status</td>
<td>tottrusted-rx.com</td>
</tr>
<tr>
<td>Order status</td>
<td>y5r.pillsfillrefill.ru</td>
</tr>
<tr>
<td>Order status</td>
<td>y5r.pillsfillrefill.ru</td>
</tr>
<tr>
<td>Order status</td>
<td>trusted-rxsi.com</td>
</tr>
<tr>
<td>Order status</td>
<td>reer.toprefilli.ru</td>
</tr>
<tr>
<td>Order status</td>
<td>cheaprx-or.com</td>
</tr>
<tr>
<td>Order status</td>
<td>trusted-rxsi.com</td>
</tr>
</tbody>
</table>

(Partial listing – 298 messages were found)
A Different “Look & Feel”
IP Check

By querying the IP address of the domain, we can see that we have seen 45 other domains hosted on the same IP address.

```
select * from domain_ip where domain = 'cheaprx-ra.com';
domain     | tld |   ip_address | start_date | last_date
----------------+-----+----------------+------------+-----------
cheaprx-ra.com | com | 91.210.64.3/32 | 2010-11-30 | 2010-11-30
(1 row)
```

```
select * from domain_ip where ip_address = '91.210.64.3/32'
order by last_date desc;
```
Another IP from the same cluster showed that 118 additional domains had been hosted on “61.187.235.250” including “pillsrefilll.ru” on December 1st.

<table>
<thead>
<tr>
<th>domain</th>
<th>tld</th>
<th>ip_address</th>
<th>start_date</th>
<th>last_date</th>
</tr>
</thead>
<tbody>
<tr>
<td>pillsrefilll.ru</td>
<td>ru</td>
<td>61.187.235.250/32</td>
<td>2010-12-01</td>
<td>2010-12-02</td>
</tr>
</tbody>
</table>

Was this two distinct customers of the same spammer? Or was it two hosting locations for the same spammer?
select * from domain_ip where domain in (select domain from domain_ip where ip_address = '61.187.235.250/32') and ip_address != '61.187.235.250/32' order by ip_address;
Phishing

PHISHING TIME
Spam vs. Phishing

• Phishing is a special category of spam. Specifically, it is that category of spam about which someone might actually give a damn care about from a forensic investigators perspective.

• The reason investigators don’t care about Spam the way they care about Phishing is that they have not been provided the tools to help them understand and describe the problem.

• We hope that the UAB Spam Data Mine is part of the solution to that problem.
Not all Criminals are the same

Without additional data, you do not know which phishing site was created by a twelve year old as a prank, and which are being run by million dollar crime syndicates.
Which Site Shall We Investigate?

Victims Per Site

If we agree that some sites have more victims than others, how could we determine this?

How will this impact our behavior?

© The University of Alabama at Birmingham, 2011
Which website caused your losses?

Just because a site captured the most userids and passwords does not mean it is responsible for the greatest financial losses.

How could we tie losses to sites?
Are Some Phishing Kits More Common Than Others?

Obviously, the answer is “Yes.”

If we agree the answer is “Yes,” how has that impacted our behavior?
Drop Email Addresses

Sites Using This "drop" Email

Would you respond differently if you knew which “drop” email addresses were used most commonly to attack your brand?

- badguy@gmail.com
- badguy@yahoo.com
- badbadboy@gmail.com
Email Aliases?

How would your activities change if you knew that all of the emails were actually the same criminal?
UAB’s Phishing process
We are currently sharing “Beta 2” of our PhishIntel tool with law enforcement and selected partners.

The goal is to help investigators make the links that we described previously in order to identify their “Big Phish”
LIVE DEMO

(please refer to video or email gar@cis.uab.edu for more information)
UAB Phishing Papers

Future Problems: More Spam

– We need DRAMATICALLY MORE SPAM.

– Alternatively, we could really use DRAMATICALLY MORE SPAMMED URLs.
We Want Your Spam!

Large-Scale Automatic Classification of Phishing Pages

Colin Whittaker
Google Inc.
cwhittak@google.com

Brian Ryner
Google Inc.
bryner@google.com

Marria Nazif
Google Inc.
marria@google.com

How do I get those!?!???

<table>
<thead>
<tr>
<th></th>
<th>Apr 16–Jul 14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total URLs Received</td>
<td>446,152,060</td>
</tr>
<tr>
<td>User Submitted URLs</td>
<td>75,048</td>
</tr>
<tr>
<td>Gmail Spam URLs</td>
<td>446,093,814</td>
</tr>
</tbody>
</table>
Future Problems: CloudFetch

- Spammers know that security researchers behave differently than people who want to buy Horny Goat Weed on the Internet.
- If one of my lab machines fetches more than five drug websites from a particular affiliate, we get “life-time IP address banned”.
- Other spammers have VERY ADVANCED IP reputation systems, some even buying the same service used by banks and online businesses. They don’t allow “TOR” nodes.
- So, what can we do to ensure that we are always able to fetch and store as evidence “what the user sees?”
- Could we build either “cloud” or “private TOR” solutions to this problem that would benefit the security community?
- If these already exist, could we get some help to get UAB connected with them?
Future Problems: “relevance” for searching

- One of our teams is building a Hadoop cluster to allow NLP queries of the raw spam data.
- How do we make sure that it’s smart enough to know NOT to include as “search terms” all the garbage in the following email:
Example Problem: NLP

If you don't see picture, [click Here]
fury hippothadee proud stands enormous say altogether mischief spies excell=ing tail tempt therewithal children arrived turnbank further come pothook la=omedon parte taste hell came camels long picked heretic fiercely thought cl=others straight furthermore letters stand wiper did stood suborners sell del=icate fact filly necklace since villages compendious jupiter octavian chan=ge distributed resisting mansion boars pense flight unfortunate theatre bio=us quintus prolific noble advanced chin frail paying whelks alum. offerings whilst hymns heads hath profession honest off observed kneeled id=olmind germany dear cross saw crows understand sport promise sermonnaires f=latl changed unhappily tongued delivery philip estate danger victory lipothy=my fabric name jaunt fact abbothawks greeks crony hogs following keeper thi=nk esperruquanchuzelubelouzerireliced antidoted cavalier sorbonne kneeled g=host returns fire camels thus choose almirods proceed quantity stick commem=orate triumphant captains suit capons closebuttock keeper coloured thence c=

Spam as Evidence
© The University of Alabama at Birmingham, 2011
(the “click here”)

Today's bestsellers

**Viagra**
Our price: **$0.70**
Viagra is an oral medication used for treating male impotence (e.g., erectile dysfunction). Viagra's advantages are e.g.,

**Viagra Soft Tabs**
Our price: **$1.25**
Viagra Soft Tabs are quick-dissolving lozenges for treating male impotence. Compared to ordinary Viagra, Viagra Soft...

**Cialis Soft Tabs**
Our price: **$1.87**
Cialis Soft Tabs (Tadalafil) are quick-dissolving tabs, used to treat male impotence. Having the shortest start time...

**Cialis**
60 pills x 20 mg + 4 free pills
$141.78

**Viagra + Cialis**
30 Viagra x 100 mg
30 Cialis x 20 mg
$176.08

**Viagra + Cialis + Levitra**
Viagra 10 pills x 100 mg
Cialis 10 pills x 20 mg
Levitra 10 pills x 20 mg
$109.99
Future Problems: Evidentiary Context

• By the time some investigator wants to investigate this spam message, how will we know what the message was about?
• The destination website will have been long gone.
• This is especially troubling with REDIRECTORS. See Tech Report: **UABCIS-TR-2010-120410: "URL Shorteners Used by Online Drug Dealers"
• (in which we identify 60+ URL shorteners, some of which are clearly created BY the spammers)
We Want To Help

Gary Warner
Director of Research in Computer Forensics
A Research Partnership between
The University of Alabama at Birmingham’s
  Department of Computer & Information Sciences
&  Department of Justice Sciences

gar@uab.edu
+1.205.422.2113

Phishing Intelligence Team
PhishIntel@cis.uab.edu

Technical Reports
http://www.cis.uab.edu/forensics/TechReports

Website:
www.cis.uab.edu/forensics/

Blog:
garwarner.blogspot.com