Spamalytics: An Empirical Analysis of Spam Marketing Conversion

Chris Kanich* Christian Kreibich† Kirill Levchenko* Brandon Enright*
Geoffrey M. Voelker* Vern Paxson† Stefan Savage*

†International Computer Science Institute
Berkeley, USA
christian@icir.org, vern@cs.berkeley.edu

*Dept. of Computer Science and Engineering
University of California, San Diego, USA
{ckanich,klevchen,voelker,savage}@cs.ucsd.edu
bmenrigh@ucsd.edu
Why SPAM?

• What are the motivations for creating and spreading SPAM?
  • Money.
  • Execute at scale
  • Be annoying
  • Stock price manipulation.
  • Propaganda.
Let’s empathize with the spammer

• What investments do spammers have to make to launch a campaign?
  • Bot infrastructure.
  • Manpower – infra maintenance and content creation.
  • Hosting provider. Specialized infrastructure – bullet proofing.

• Why is SPAM delivery hard?
  • Filtering Blacklist: MTA
  • User ignore SPAM
  • Finding target audience.
  • Anti-SPAM heuristics
  • Honey-pot clients
  • Infrastructure up-time
  • Business model.
Storm botnet

- Peer-to-peer botnet that propagates through spam.
- It uses 2 kinds of protocols to communicate:
  - UDP-based Overnet protocol
  - TCP based Command and Control (C&C) protocol
- Overnet Protocol:
  - 4 basic messages- Connect, Search, Publicize, Publish.
  - Process:
    1. Choose an OID pseudo-randomly (128 bit).
    2. Connect to all the peers in bootstrap list.
    3. Each peer returns a list of 20 peers. This process is repeated until the bot has gathered enough information.
    4. Publicize itself to other peers to and search its own ID to keep up-to-date with ID churn. ---- Does this mean there is dynamic ID allocation??
    5. Publish and Search messages are used to (key, value) pair interface. The key is used to encode a dynamic code that allows bots to find each other on demand.
    6. These keys are used to locate and communicate with Storm C&C nodes. If a bot wishes to become a C&C node, it can use the time base hashing algorithm to encode the key and its own ip and port can be encoded in the value.

Why proxy bots?
- Worker may not be publicly accessible.
- Hiding master id/spammer.
- Bandwidth.
- Multiple actors involved.
Storm botnet

- Worker bots make request for workload by connecting to proxy servers. (Bots initiate communication: “Pull based”)
- Proxy servers connect request from bots to master servers.
- Master servers send Commands to bots and receive reports.
- Bots send spam according to commands received from master server.
- Role of an instance, whether its going to be a worker bot or a proxy, is decided dynamically:
  - If an infected system can be accessed externally, it is eligible to become a proxy or else it becomes a worker bot.
- Workload from master has 3 components:
  1. Spam template(s). Written in custom macro language used for generating “valid” emails.
  2. Delivery list of emails.
  3. Named dictionaries containing target email addresses, subject lines etc.
Methodology

• Create hosts infected by Storm and make them externally accessible -> proxy bots on demand.
• These hosts were instantiated within a virtual environment.
• Traffic routed through controlled gateway to monitor a traffic.
• Traffic is intercepted and C&C commands are re-written.
• Target sites, spam templates and delivery list are modified as part of the re-writing process.
• A click-based network element was created to re-direct C&C traffic incoming to the gateway to a SOCKS proxy that re-writes the C&C message and sends the msg to the click element.
• Finally the click element forwards the modified C&C message to intended proxy bot.
• When a worker bot sends success report for a delivery to modified target email address, the corresponding report is removed before sending it back to master server.
Methodology – Discussion

• Merits:
  • Good Control metrics

• Demerits:
  • Ethical nature.
  • Very tethered to existing storm campaign (geographic region).
  • Complexity, Scalability.
  • Heuristically estimation, lack GT.
Measurements

• Email Delivery:
  • Add emails in delivery list: gmail, yahoo, hotmail, dept. email w/ Barracuda spam filtering, SMTP “sinks” that accept all mails sent to them.

• Click Through and Conversion: modifying target site url.
  • Self-Propagation:
    • Download “benign” malware executable.
    • If this benign malware is executed, it sends a HTTP POST to a destination.
    • Access to download link is logged and can be correlated to HTTP POST to calculate “conversion”.
  • Advertisement:
    • Sale of pharmaceutical products. Users can add add product to cart but cannot checkout (404 Error).
    • Every spam mail has a unique identifier to associate site visit to spam email. (this feature not available for self-propagation)

• Separating Crawlers: Create a global IP blacklist
  • Ad website access w/o unique identifier.
  • Access robots.txt (accessed by legitimate crawlers)
  • Malformed requests
  • Disabled javascript and do not load images.
  • Multiple access to Ad website with same UID but different User-Agent field.
  • Multiple malware binary download (>10).
  • Introduce target IPs in self-propagation dictionary during “inactive” period.
Measurements - Discussion

- Scope of error in conversion estimation?
  - Estimating Crawlers might be wrong.
  - Checkout is not “sale”.
  - Region-based targeting not considered.

- What kind of access can still impact the measurement?
Measurements – Ethics Discussion

• Ethos: “We strictly reduce harm”

• Features supporting this principle:
  • No additional harm.

• Possibility of harm/vulnerability:
  • Executable download w/o consent
  • Spurious logic
  • Collateral damage, retaliation.
Experimental Results – Campaign datasets

- 3 Campaigns observed: Postcard, April Fools, Pharmacy
- Number of workers connected to a proxy is constant.
- Most workers connect to proxy once -> access diversity
- Most workers do not connect to multiple proxies.
- On avg. workers remain connect to a proxy for 40 mins.

Thoughts on presented statistics:

Table 1: Campaigns used in the experiment.

<table>
<thead>
<tr>
<th>Campaign</th>
<th>Dates</th>
<th>Workers</th>
<th>E-mails</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pharmacy</td>
<td>Mar 21 – Apr 15</td>
<td>31,348</td>
<td>347,590,389</td>
</tr>
<tr>
<td>Postcard</td>
<td>Mar 9 – Mar 15</td>
<td>17,639</td>
<td>83,865,479</td>
</tr>
<tr>
<td>April Fool</td>
<td>Mar 31 – Apr 2</td>
<td>3,678</td>
<td>38,651,124</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>469,906,992</strong></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: The 10 most-targeted e-mail address domains and their frequency in the combined lists of targeted addresses over all three campaigns.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>hotmail.com</td>
<td>8.47%</td>
</tr>
<tr>
<td>yahoo.com</td>
<td>5.05%</td>
</tr>
<tr>
<td>gmail.com</td>
<td>3.17%</td>
</tr>
<tr>
<td>aol.com</td>
<td>2.33%</td>
</tr>
<tr>
<td>yahoo.co.in</td>
<td>1.13%</td>
</tr>
<tr>
<td>sbglobal.net</td>
<td>0.93%</td>
</tr>
<tr>
<td>mail.ru</td>
<td>0.86%</td>
</tr>
<tr>
<td>shaw.ca</td>
<td>0.61%</td>
</tr>
<tr>
<td>waradoo.fr</td>
<td>0.61%</td>
</tr>
<tr>
<td>msn.com</td>
<td>0.58%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>25.19%</td>
</tr>
</tbody>
</table>
Experimental Results – Spam Conversion Pipeline

• Stage A-B: invalid address, Mail Transfer Agent (MTA) blacklisting.
• MTA delivery stats depends on worker bot response.
• Spam classification by different services was measured using some test emails created by authors.
• Observation:
  • Most emails get rejected at MTA level.
  • Public email services do pretty good job at detecting spam.

Thoughts on this procedure:

<table>
<thead>
<tr>
<th>STAGE</th>
<th>PHARMACY</th>
<th>POSTCARD</th>
<th>APRIL FOOL</th>
</tr>
</thead>
<tbody>
<tr>
<td>A – Spam Targets</td>
<td>347,590,389</td>
<td>82,700,000</td>
<td>40,155,487</td>
</tr>
<tr>
<td>B – MTA Delivery (est.)</td>
<td>82,700,000</td>
<td>21,100,000</td>
<td>10,155,487</td>
</tr>
<tr>
<td>C – Inbox Delivery</td>
<td>10,522</td>
<td>3,827</td>
<td>2,721</td>
</tr>
<tr>
<td>D – User Site Visits</td>
<td>28</td>
<td>316</td>
<td>225</td>
</tr>
<tr>
<td>E – User Conversions</td>
<td>0.00030%</td>
<td>0.00457%</td>
<td>0.00680%</td>
</tr>
</tbody>
</table>

Table 3: Filtering at each stage of the spam conversion pipeline for the self-propagation and pharmacy campaigns. Percent to the conversion rate relative to Stage A.

<table>
<thead>
<tr>
<th>SPAM FILTER</th>
<th>PHARMACY</th>
<th>POSTCARD</th>
<th>APRIL FOOL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gmail</td>
<td>0.00683%</td>
<td>0.00176%</td>
<td>0.00226%</td>
</tr>
<tr>
<td>Yahoo</td>
<td>0.00173%</td>
<td>0.000542%</td>
<td>none</td>
</tr>
<tr>
<td>Hotmail</td>
<td>none</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>Barracuda</td>
<td>0.131%</td>
<td>N/A</td>
<td>0.00826%</td>
</tr>
</tbody>
</table>

Table 4: Number of messages delivered to a user’s inbox as a fraction of those injected for test accounts at free e-mail providers and a commercial spam filtering appliance. The test account for the Barracuda appliance was not included in the Postcard campaign.
Experimental Results – Time to Click

- How long the spam infrastructure must be active to catch all potential sales?
- Crawlers visit the site almost instantly after the spam is sent.
- This experiment was not done on self-propagating site as there is no UID to link mail delivery to site visit.

Thoughts about this experiment:

Figure 7: Time-to-click distributions for accesses to the pharmacy site.
Blacklisting

- Composite Black List (CBL) list of host addresses which have been known to send spam.
- CBL of the IT department of the author’s university was monitored.
- 81% of workers got blacklisted.
- 77% of blacklisted were already blacklisted prior to observation (from previous campaigns).
- Most of the new workers got on blacklist within 1.5 hours of sending spam.
- Moral of the story: “Don’t send spam don’t get blacklisted.”
- Most workers experience 5 cycles of listing followed by delisting.

Thoughts about this procedure:
Conversion Analysis

- Geographic Location of hosts that “convert”.
- Nation based response analysis:
  - Response Rate: Stage A-D (not E)
  - Country specific TLDs were considered. Generic TLD like .com were ignored. Gmail, yahoo, Hotmail etc.??
  - Per country response rate dependent on structural causes: spam filtering, anti-spam education etc.

Does this statistics give any economic insight? (objective of this work)
Facebook's average revenue per user as of 4th quarter 2019, by region

*(in U.S. dollars)*

<table>
<thead>
<tr>
<th></th>
<th>Worldwide</th>
<th>U.S. and Canada</th>
<th>Europe</th>
<th>Asia Pacific</th>
<th>Rest of world</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q4 '11</td>
<td>1.38</td>
<td>3.2</td>
<td>1.6</td>
<td>0.56</td>
<td>0.41</td>
</tr>
<tr>
<td>Q4 '12</td>
<td>1.54</td>
<td>4.08</td>
<td>1.71</td>
<td>0.69</td>
<td>0.56</td>
</tr>
<tr>
<td>Q4 '13</td>
<td>2.14</td>
<td>6.03</td>
<td>2.61</td>
<td>0.95</td>
<td>0.84</td>
</tr>
<tr>
<td>Q4 '14</td>
<td>2.81</td>
<td>9</td>
<td>3.45</td>
<td>1.27</td>
<td>0.94</td>
</tr>
<tr>
<td>Q4 '15</td>
<td>3.73</td>
<td>13.7</td>
<td>4.56</td>
<td>1.6</td>
<td>1.1</td>
</tr>
<tr>
<td>Q4 '16</td>
<td>4.83</td>
<td>19.81</td>
<td>5.98</td>
<td>2.07</td>
<td>1.41</td>
</tr>
<tr>
<td>Q4 '17</td>
<td>6.18</td>
<td>26.76</td>
<td>8.86</td>
<td>2.54</td>
<td>1.86</td>
</tr>
<tr>
<td>Q4 '18</td>
<td>7.37</td>
<td>34.86</td>
<td>10.98</td>
<td>2.96</td>
<td>2.11</td>
</tr>
<tr>
<td>Q4 '19</td>
<td>8.52</td>
<td>41.41</td>
<td>13.21</td>
<td>3.57</td>
<td>2.48</td>
</tr>
</tbody>
</table>

Conclusion

• 26 days, 350 M emails => 28 “sales”.
• Storm pharmaceutical spam revenue: $3.5M/year. (estimate)
• Storm spam revenue = $1.75M/year
• Email cost: $80/M. Total = $25,000. (For 26 days)
• Authors suggest this price is too high, and the business model makes sense if operators of the botnet and affiliates selling the products are one and the same.
• Evidence supporting above claim: email addresses appearing in self-propagating campaign eventually end up in pharmacy campaign.
• Exact labor cost cannot be calculated as it depends on region of development.