Spam: More than Just a Nuisance

Spam: unsolicited bulk emails
Ham: legitimate emails from desired contacts

- 95% of all email traffic is spam
  (Sources: Microsoft security report, MAAWG and Spamhaus)
  - In 2009, the estimation of lost productivity costs is $130 billion worldwide
    (Source: Ferris Research)

- Spam is the carrier of other attacks
  - Phishing
  - Virus, Trojan horses, …

by S. Hao, N. A. Syed, N. Feamster, A. Gray, S. Krasser
Current Anti-spam Methods

- **Content-based filtering:** *What is in the mail?*
  - More spam format rather than text (PDF spam ~12%)
  - Customized emails are easy to generate
  - High cost to filter maintainers

- **IP blacklist:** *Who is the sender?* (e.g., DNSBL)
  - ~10% of spam senders are from previously unseen IP addresses (due to dynamic addressing, new infection)
  - ~20% of spam received at a spam trap is not listed in any blacklists
SNARE: Our Idea

- Spatio-temporal Network-level Automatic Reputation Engine
  - Network-Based Filtering: How the email is sent?
    - Fact: > 75% spam can be attributed to botnets
    - Intuition: Sending patterns should look different than legitimate mail
  - Example features: geographic distance, neighborhood density in IP space, hosting ISP (AS number) etc.
  - Automatically determine an email sender’s reputation
    - 70% detection rate for a 0.2% false positive rate
Why Network-Level Features?

- Lightweight
  - Do not require content parsing
    - Even getting one single packet
    - Need little collaboration across a large number of domains
  - Can be applied at high-speed networks
  - Can be done anywhere in the middle of the network
    - Before reaching the mail servers
- More Robust
  - More difficult to change than content
  - More stable than IP assignment
Talk Outline

• Motivation
• **Data From McAfee**
• Network-level Features
• Building a Classifier
• Evaluation
• Future Work
• Conclusion
Data Source

- McAfee’s TrustedSource email sender reputation system
  - Time period: 14 days October 22 – November 4, 2007
  - Message volume: Each day, 25 million email messages from 1.3 million IPs
  - Reported appliances 2,500 distinct appliances (≈ recipient domains)
  - Reputation score: certain ham, likely ham, certain spam, likely spam, uncertain
Finding the Right Features

• Question: Can sender reputation be established from just a single packet, plus auxiliary information?
  – Low overhead
  – Fast classification
  – In-network
  – Perhaps more evasion resistant

• Key challenge
  – What features satisfy these properties and can distinguish spammers from legitimate senders?
Network-level Features

• Feature categories
  – Single-packet features
  – Single-header and single-message features
  – Aggregate features

• A combination of features to build a classifier
  – No single feature needs to be perfectly discriminative between spam and ham

• Measurement study
  – McAfee’s data, October 22-28, 2007 (7 days)
### Summary of SNARE Features

<table>
<thead>
<tr>
<th>Category</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single-packet</strong></td>
<td>geodesic distance between the sender and the recipient</td>
</tr>
<tr>
<td></td>
<td>average distance to the 20 nearest IP neighbors of the sender</td>
</tr>
<tr>
<td></td>
<td>probability ratio of spam to ham when getting the message</td>
</tr>
<tr>
<td></td>
<td>status of email-service ports on the sender</td>
</tr>
<tr>
<td></td>
<td>AS number of the sender’s IP</td>
</tr>
<tr>
<td><strong>Single-header/message</strong></td>
<td>number of recipient</td>
</tr>
<tr>
<td></td>
<td>length of message body</td>
</tr>
<tr>
<td><strong>Aggregate features</strong></td>
<td>average of message length in previous 24 hours</td>
</tr>
<tr>
<td></td>
<td>standard deviation of message length in previous 24 hours</td>
</tr>
<tr>
<td></td>
<td>average recipient number in previous 24 hours</td>
</tr>
<tr>
<td></td>
<td>standard deviation of recipient number in previous 24 hours</td>
</tr>
<tr>
<td></td>
<td>average geodesic distance in previous 24 hours</td>
</tr>
<tr>
<td></td>
<td>standard deviation of geodesic distance in previous 24 hours</td>
</tr>
</tbody>
</table>

**Total of 13 features in use**
What Is In a Packet?

• Packet format (incoming SMTP example)

- IP Header
  - Source IP, Destination IP

- TCP Header
  - Destination port: 25

- SMTP
  - Text Command
    - Empty for the first packet

• Help of auxiliary knowledge:
  - Timestamp: the time at which the email was received
  - Routing information
  - Sending history from neighbor IPs of the email sender
Sender-receiver Geodesic Distance

- **Intuition:**
  - Social structure limits the region of contacts
  - The geographic distance travelled by spam from bots is close to random

by S. Hao, N. A. Syed, N. Feamster, A. Gray, S. Krasser
Distribution of Geodesic Distance

- Find the physical latitude and longitude of IPs based on the MaxMind’s GeoIP database
- Calculate the distance along the surface of the earth

90% of legitimate messages travel 2,500 miles or less

- Observation: Spam travels further

by S. Hao, N. A. Syed, N. Feamster, A. Gray, S. Krasser
Sender IP Neighborhood Density

• Intuition:
  – The infected IP addresses in a botnet are close to one another in numerical space
  – Often even within the same subnet
Distribution of Distance in IP Space

- IPs as one-dimensional space (0 to $2^{32}-1$ for IPv4)
- Measure of email sender density: the average distance to its $k$ nearest neighbors (in the past history)

For spammers, $k$ nearest senders are much closer in IP space

- Observation: Spammers are surrounded by other spammers

by S. Hao, N. A. Syed, N. Feamster, A. Gray, S. Krasser
• Intuition:
  – Diurnal sending pattern of different senders
  – Legitimate email sending patterns may more closely track workday cycles
Differences in Diurnal Sending Patterns

- Local time at the sender’s physical location
- Relative percentages of messages at different time of the day (hourly)

Observation: Spammers send messages according to machine power cycles
Status of Service Ports

- Ports supported by email service provider

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Port</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMTP</td>
<td>25</td>
</tr>
<tr>
<td>SSL SMTP</td>
<td>465</td>
</tr>
<tr>
<td>HTTP</td>
<td>80</td>
</tr>
<tr>
<td>HTTPS</td>
<td>443</td>
</tr>
</tbody>
</table>

- Intuition:
  - Legitimate email is sent from other domains’ MSA (Mail Submission Agent)
  - Bots send spam directly to victim domains
Distribution of number of Open Ports

- Actively probe back senders’ IP to check out what service ports open
- Sampled IPs for test, October 2008 and January 2009

90% of spamming IPs have none of the standard mail service ports open

Observation: Legitimate mail tends to originate from machines with open ports

by S. Hao, N. A. Syed, N. Feamster, A. Gray, S. Krasser
AS of sender’s IP

- Intuition: Some ISPs may host more spammers than others

- Observation: A significant portion of spammers come from a relatively small collection of ASes*
  - More than 10% of unique spamming IPs originate from only 3 ASes
  - The top 20 ASes host ~42% of spamming IPs

# Summary of SNARE Features

<table>
<thead>
<tr>
<th>Category</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-packet</td>
<td>geodesic distance between the sender and the recipient</td>
</tr>
<tr>
<td></td>
<td>average distance to the 20 nearest IP neighbors of the sender</td>
</tr>
<tr>
<td></td>
<td>probability ratio of spam to ham when getting the message</td>
</tr>
<tr>
<td></td>
<td>status of email-service ports on the sender</td>
</tr>
<tr>
<td></td>
<td>AS number of the sender’s IP</td>
</tr>
<tr>
<td>Single-header/message</td>
<td>number of recipient</td>
</tr>
<tr>
<td></td>
<td>length of message body</td>
</tr>
<tr>
<td>Aggregate features</td>
<td>average of message length in previous 24 hours</td>
</tr>
<tr>
<td></td>
<td>standard deviation of message length in previous 24 hours</td>
</tr>
<tr>
<td></td>
<td>average recipient number in previous 24 hours</td>
</tr>
<tr>
<td></td>
<td>standard deviation of recipient number in previous 24 hours</td>
</tr>
<tr>
<td></td>
<td>average geodesic distance in previous 24 hours</td>
</tr>
<tr>
<td></td>
<td>standard deviation of geodesic distance in previous 24 hours</td>
</tr>
</tbody>
</table>

**Total 13 features in use**
RuleFit (ensemble learning)

- $F(x) = \alpha_0 + \sum_{m=1}^{M} a_m f_m(x)$
- $F(x)$ is the prediction result (label score)
- $f_m(x)$ are base learners (usually simple rules)
- $a_m$ are linear coefficients

Example

<table>
<thead>
<tr>
<th>Rule</th>
<th>$F(x)$</th>
<th>$a_m$</th>
<th>$f_m(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule 1</td>
<td>0.080</td>
<td>0.080</td>
<td>Geodesic distance &gt; 63 AND AS in (1901, 1453, …)</td>
</tr>
<tr>
<td>Rule 2</td>
<td>+0</td>
<td>0.257</td>
<td>Port status: no SMTP service listening</td>
</tr>
</tbody>
</table>

Feature instance of a message

Geodesic distance = 92, AS=1901, port SMTP is open
Talk Outline

• Motivation
• Data From McAfee
• Network-level Features
• Building a Classifier

• **Evaluation**
  – Setup
  – Accuracy
  – Detecting “Fresh” Spammers
  – In Paper: Retraining, Whitelisting, Feature Correlation

• Future Work
• Conclusion
Evaluation Setup

- **Data**
  - 14-day data, October 22 to November 4, 2007
  - 1 million messages sampled each day (only consider certain spam and certain ham)

- **Training**
  - Train SNARE classifier with equal amount of spam and ham (30,000 in each categories per day)

- **Temporal Cross-validation**
  - Temporal window shifting

---

![Evaluation Diagram](image-url)

"Trial 1  Trial 2"

---

"Train  Test"

Data subset

---

by S. Hao, N. A. Syed, N. Feamster, A. Gray, S. Krasser
Receiver Operator Characteristic (ROC)

- False positive rate = Misclassified ham/Actual ham
- Detection rate = Detected spam/Actual spam
  (True positive rate)

FP under detection rate 70%

<table>
<thead>
<tr>
<th>False Positive</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Packet</td>
<td>0.44%</td>
</tr>
<tr>
<td>Single Header/Message</td>
<td>0.29%</td>
</tr>
<tr>
<td>24+ Hour History</td>
<td>0.20%</td>
</tr>
</tbody>
</table>

As a first of line of defense, SNARE is effective.
Detection of “Fresh” Spammers

• “Fresh” senders
  – IP addresses not appearing in the previous training windows
• Accuracy
  – Fixing the detection rate as 70%, the false positive is 5.2%

SNARE is capable of automatically classifying ‘fresh’ spammers (compared with DNSBL)

by S. Hao, N. A. Syed, N. Feamster, A. Gray, S. Krasser
Future Work

• Combine SNARE with other anti-spam techniques to get better performance
  – Can SNARE capture spam undetected by other methods (e.g., content-based filter)?

• Make SNARE more evasion-resistant
  – Can SNARE still work well under the intentional evasion of spammers?
Conclusion

- Network-level features are effective to distinguish spammers from legitimate senders
  - Lightweight: Sometimes even by the observation from one single packet
  - More Robust: Spammers might be hard to change all the patterns, particularly without somewhat reducing the effectiveness of the spamming botnets

- SNARE is designed to automatically detect spammers
  - A good first line of defense