Motivations

• Huge volume of bot queries today
  – Finding security vulnerabilities [Usenix Sec 2010]
  – Search engine optimizations [Usenix Sec 2011]
  – … …

• Understanding bot queries can be a new direction to tackle network security
  – For security analysts: identify attack trends and detect botnets
  – For search engines: throttle malicious activities and protect good users
Related Work

F. Yu, et al., “Large-Scale Search Bot Detection”, WSDM 2010
......

......

Search Queries → Bot Query Detection → Bot Queries

User Queries → Query Classification and Clustering → Auto Suggestion

Bot Queries → ?
Our Work: SBotScope

- What are bot queries searching for?
- Who submitted bot queries? Botnets? Data centers?

- Search Queries
- Bot Queries
- User Queries
- Bot Query Detection
- Query Classification and Clustering

SBotScope
Challenges

• Different behaviors compared to real users
  – Fast-changing topics
  – Few bot queries have clicks

• Various obfuscation strategies
  – Mix truly intended queries with legitimate ones
  – Add stuffing words to each query

• Huge complexity
  – Large data volume
  – No data labeling
Query Pattern Generation

- **Word Combinations**
  - Robust to word order changes

- **Pattern**: a *specific* word-combination that happens *frequently*
  - Frequent: frequency rankings
  - Specific: conditional entropy + normal query popularity
Topic Analysis

• **Syntactically:** construct pattern trees hierarchically

• **Semantically:** group trees into topics via spectral clustering

![Diagram showing hierarchical tree structure with categories and subcategories related to pre-owned cars.]
Feature Vector Generation and Host Clustering

Host1  
Query1  
Query2  
.....  
Pattern1  
Pattern2  
Pattern3

Host2  
Query  
.....  
Pattern4  
Pattern5

Host3  
Query  
.....  
Pattern1  
Pattern2

Host4  
Query  
.....  
Pattern1  
Pattern2

Pattern Tree  
Match  
Clustering
Host Cluster Classification

![Diagram showing data points classified into likely data centers and likely botnets based on IP activity and prefix distribution.]

- Likely data centers
- Likely botnets
Results: Complexity Reduction

- Normal Search Queries
  - Query Pattern Generation: 726K query patterns
  - Pattern Tree Construction: 1,823 pattern trees
  - Topic Clustering: 50 topics

- Bot Search Queries
  - Host Representation and Feature Vector Generation: 14 million vectors
  - Host Clustering and Classification: 37K botnet clusters, 137 data center clusters

- 3 billion bot queries
## Intentions: Popular Topics

<table>
<thead>
<tr>
<th>Topic</th>
<th>% of Bot Queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vulnerability Discovery</td>
<td>32.8%</td>
</tr>
<tr>
<td>Email Harvest</td>
<td>11%</td>
</tr>
<tr>
<td>Content Download</td>
<td>3.6%</td>
</tr>
<tr>
<td>Fashion Items</td>
<td>1.4%</td>
</tr>
<tr>
<td>Car Sale</td>
<td>1.3%</td>
</tr>
<tr>
<td>News</td>
<td>0.7%</td>
</tr>
<tr>
<td>......</td>
<td>......</td>
</tr>
</tbody>
</table>

Top 6 popular topics
## Intentions: Popular Topics

<table>
<thead>
<tr>
<th>Topic</th>
<th>% of Bot Queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vulnerability Discovery</td>
<td>32.8%</td>
</tr>
<tr>
<td>Email Harvest</td>
<td>11%</td>
</tr>
<tr>
<td>Content Download</td>
<td>3.6%</td>
</tr>
<tr>
<td>Fashion Items</td>
<td></td>
</tr>
<tr>
<td>Car Sale</td>
<td></td>
</tr>
<tr>
<td>News</td>
<td></td>
</tr>
<tr>
<td>......</td>
<td></td>
</tr>
</tbody>
</table>

### Top 5 Patterns

- list members mode php
- mode php register
- es php page
- aspx html php
- powered by

### Query Patterns

<table>
<thead>
<tr>
<th>Query Patterns</th>
<th>Vulnerabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>powered by photo album</td>
<td>PHP Album 0.3.2.3 Remote command execution</td>
</tr>
<tr>
<td>powered by update</td>
<td>PHP-Update 2.7 Remote code execution</td>
</tr>
<tr>
<td>Powered by icalendar</td>
<td>PHP-iCalendar 2.24 File upload</td>
</tr>
</tbody>
</table>
## Origination: Botnets vs. Data Centers

<table>
<thead>
<tr>
<th></th>
<th>Botnet clusters</th>
<th>Data center clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number</strong></td>
<td>37,268</td>
<td>137</td>
</tr>
<tr>
<td><strong>Intentions</strong></td>
<td>Vulnerability, email address</td>
<td>Commercially relevant information</td>
</tr>
<tr>
<td><strong>Queries/host/day</strong></td>
<td>17</td>
<td>475</td>
</tr>
<tr>
<td><strong>Network distributions</strong></td>
<td>Different organizations</td>
<td>The same organizations</td>
</tr>
<tr>
<td><strong>Top country</strong></td>
<td>CN, MX, IT, BR</td>
<td>US</td>
</tr>
</tbody>
</table>
Case Study I: A Botnet Cluster

```
“index.php/thread-” shooter
“yabb/yabb.pl?board” siver
“index.php/thread-” site:com

......

act “WordPress forum plugin by Fredrik Fahlstad” site:.edu
```

```
inch “WordPress forum plugin by Fredrik Fahlstad” site:.cn plus
“WordPress forum plugin by Fredrik Fahlstad” site:.net change
“WordPress forum plugin by Fredrik Fahlstad” site:.com
```
Case Study II: A Data Center Cluster

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td># of IPs</td>
<td>446</td>
</tr>
<tr>
<td># of /24 prefixes</td>
<td>29</td>
</tr>
<tr>
<td># of organizations</td>
<td>1</td>
</tr>
<tr>
<td># of queries</td>
<td>873,064</td>
</tr>
</tbody>
</table>

“alice foo” + (“alice is” OR “alice was”) –Genealogy – Generation language:en
Case Study III: Vulnerability-Searching Data Centers

<table>
<thead>
<tr>
<th></th>
<th># of IPs</th>
<th># of organizations</th>
<th># of queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>A cluster in May</td>
<td>41</td>
<td>1</td>
<td>244,054</td>
</tr>
<tr>
<td>A cluster in Oct</td>
<td>18</td>
<td>1</td>
<td>113,994</td>
</tr>
</tbody>
</table>

Query Pattern: “powered by”
Conclusion

• SBotScope allows systematic analysis of bot queries
  – Understand query intentions and origination
  – Increase detection coverage
  – Fully automated and scalable

• A new direction to improve network security
  – Identify attack trends at their early stages
  – Detect botnets and malicious data center hosts
Thank you!

- Questions ...