Adaptive Attacks and Defenses in Denial of Information

Calton Pu (Georgia Tech), Kang Li and L. Ramaswamy (U. Georgia)

http://www-static.cc.gatech.edu/projects/doi  NSF Grant CNS-0716484 & 0716357

Problem Impact and Challenge

- Email spam flood: [MAAWG 2006Q1] 370 Billion messages (80%) blocked or tagged by the backbone; typical servers filter another 80%; typical email clients filter 80% more.
- Spam/phishing/deception a pervasive problem in web pages, blogosphere, social networks, other media.

Adversarial Learning Challenge

1. Attacks evolve, e.g., email camouflage attacks by randomized good words
2. Defenses adapt and learn to identify new attacks
3. But attacks provide deceptive input to mislead the automated learning defense

Systematic and Robust Defenses

- Hypothesis: Automated generation of spam also creates identifiable patterns of deception
- Automated detection of such patterns of deception in many communications media
- Experimental evaluation based on large, real data sets

<table>
<thead>
<tr>
<th>Method</th>
<th>Patterns of Deception</th>
<th>Large Real Dataset for Evaluation</th>
<th>Research and Practical Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Media</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Email spam</td>
<td>Presence of strong</td>
<td>Spam archive and Enron data set</td>
<td>Effective defense w/o retraining</td>
</tr>
<tr>
<td>(camouflage attack)</td>
<td>spam tokens</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image spam</td>
<td>Presence of strong</td>
<td>Spam archive, TREC 2005</td>
<td>Effective detection of image spam</td>
</tr>
<tr>
<td>(content analysis)</td>
<td>colors and large text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Web spam</td>
<td>Links to known spam</td>
<td>WebBase, Spam Corpus</td>
<td>Effective prediction of web spam pages</td>
</tr>
<tr>
<td>(credit-based link analysis)</td>
<td>sites (or nearby sites)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Web spam (HTTP header analysis)</td>
<td>Web host IP, header pref. of spam pages</td>
<td>WebbSpam, WebBase, UK2006</td>
<td>Effective prediction of spam before upload</td>
</tr>
<tr>
<td>Privacy-aware collaborative filters</td>
<td>Fine-grained similarity in text</td>
<td>TREC, Spam Assas. Corpus</td>
<td>Effective identification of spam w/ privacy</td>
</tr>
</tbody>
</table>

Result 1: Email Camouflage Defense

Low false positives and low false negatives with 25 legitimate and 9000 spam features

Result 2: Image Spam Analysis

Simple Colors

Self-Similar

What’s common to the above image spam?

Structural patterns of deception:
- Color heterogeneity
- Color conspicuousness
- Self-similarity when subdivided

Result 3: Privacy-Aware Collaborative Spam Filtering

Cyber Trust Principal Investigators Meeting
March 16-18, 2008; New Haven, CT