Towards Measuring the Effectiveness of Telephony Blacklists

Sharbani Pandit (Georgia Tech), Roberto Perdisci (UGA, Georgia Tech), Mustaque Ahamad (Georgia Tech), Payas Gupta (Pindrop)
Internet Threats Are Moving to Telephony

- The phone voice channel is largely unprotected (phone spam via robocalls, phishing etc.).
- 3M+ online complaints regarding unsolicited calls were reported in 2017.
- Call blocking applications (True Caller, YouMail etc.) have emerged.
- Public information not available.
Research Goal

Systematic investigation of multiple data sources that may be leveraged to automatically learn phone blacklists, and explore the potential effectiveness of such blacklists by indirectly measuring their ability to block future unwanted phone calls.

Data Sources

Evaluation
Phone Abuse Data Sources

FTC

COC

CDR

HCT

Context-less

Context-rich
Contributions

● First systematic study of estimating the effectiveness of phone blacklists.

● We investigate a number of alternative approaches for building phone blacklists.

● Blacklists are capable of blocking a significant fraction of future unwanted calls (e.g., 55% or more of unsolicited calls).

● We use a combination of unsupervised learning techniques to discover campaigns.
<table>
<thead>
<tr>
<th></th>
<th>Dataset Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>FTC</td>
<td>Number of FTC reports</td>
</tr>
<tr>
<td></td>
<td>Number of Callers</td>
</tr>
<tr>
<td>CDR</td>
<td>Number of CDRs</td>
</tr>
<tr>
<td></td>
<td>Number of Callers</td>
</tr>
<tr>
<td></td>
<td>Number of Callees</td>
</tr>
<tr>
<td>COC</td>
<td>Number of User Comments</td>
</tr>
<tr>
<td></td>
<td>Number of Callers</td>
</tr>
<tr>
<td>HCT</td>
<td>Number of Transcripts</td>
</tr>
<tr>
<td></td>
<td>Number of Callers</td>
</tr>
<tr>
<td></td>
<td>Number of Callees</td>
</tr>
</tbody>
</table>
Ground Truth

● Context (content of the call) when available
● YouMail, TrueCaller
● Whitepages
● We collect legitimate numbers from Yellowpages
● Calling back phone numbers
Blacklisting System Overview

Data Collection
- HCT
- COC
- COCNC
- CDR
- FTC

Data Pre-processing
- Remove Stop Words → Bag of words → tf-idf
  - Reduce Noise
- Filter Out Low Volume Phone Numbers

Blacklist Formation
- LSI Topic Modeling → Calculate Topic Score
- LSI Topic Modeling → Calculate Topic Score
- Add Remaining Phone Numbers
- One Class Learning
- Add Remaining Phone Numbers

Extract Phone Numbers
- HCT Blacklist
- COC Blacklist
- COCNC Blacklist
- CDR Blacklist
- FTC Blacklist
Context-less Blacklisting

Blacklisting using the CDR data:

- Calculate blacklist score for each phone number $p_i$
  \[ s(p_i, \Delta t) = \alpha \times \text{vol}(p_i, \Delta t) + \beta \times \text{nod}(p_i, \Delta t) \]
  
  Any number $p_i$ whose blacklist score is greater than a predetermined threshold $\theta_b$ is added to the blacklist.

Blacklisting using the FTC/COCNC data:

- To filter out this possible noise, we exclude all phone numbers that have been reported in less than $\theta_c$ complaints.
- All remaining numbers are then simply added to the blacklist.
Context-rich Blacklisting

Blacklisting using the HCT and COC data:

1. HCT/COC
2. Filter out noise
3. Bag of words and tf-idf
4. LSI topic modeling
5. Label topics
6. Group calls belonging to spam campaigns
7. Blacklist of phone numbers
# Topic modeling results

<table>
<thead>
<tr>
<th>Topic</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Listing</td>
<td>google, listing, front page, business, verify, press, removed, search, locally</td>
</tr>
<tr>
<td>Free Cruise</td>
<td>cruise, survey, bahamas, awarded, correctly, included, participate, congratulation</td>
</tr>
<tr>
<td>Google listing</td>
<td>listing, verify, front, google, page, updated, record, show, end, list</td>
</tr>
<tr>
<td>Business Verification</td>
<td>verification, address, name, phone, number, cancel, flagged, map, notice, business</td>
</tr>
<tr>
<td>Topic 4</td>
<td>hotel, pressed, exclusive, telephone, husband, marriott, detail, announcement</td>
</tr>
<tr>
<td>Topic 5</td>
<td>hotel, exclusive, husband, marriott, star, stay, placed, complimentary, further</td>
</tr>
<tr>
<td>Topic 6</td>
<td>electricity, bill, per, system, stop, increase, energy, renewable, soon, coming</td>
</tr>
<tr>
<td>Topic 7</td>
<td>optimize, found, date, order, indicate, critical, online, updated, show, end</td>
</tr>
<tr>
<td>Topic 8</td>
<td>system, interest, eligibility, cost, account, rate, credit, notice, card, lower</td>
</tr>
</tbody>
</table>
Results
Blacklist size and overlap
Call Blocking Rates
Call Blocking Rates

Free Cruise

Tech Support
Comparison with Third-party BLs

- A random sample of 12500 phone numbers were taken from our datasets.
- 2.4% were labeled as spam by Youmail.
- We found 87% of these phone numbers in one or more of our blacklists.

- We used Whitepages to perform reverse lookup for the blacklisted numbers.
- Most of the numbers are VoIP numbers.
- Most of the numbers do not have owner information.
False Positives

- We crawled 100,000 benign phone numbers of businesses randomly chosen from Yellow- Pages.
- We found a false positive rate of 0.01%.
Caller ID spoofing

- Caller ID spoofing causes call blocking rates to drop.
- Not all campaigns use caller ID spoofing aggressively.
- Recent initiatives that address caller ID spoofing
  - FCC rules to block calls from unassigned phone numbers.
  - Industry led “Strikeforce” - Stir, Shaken
Conclusion

- Systematic study of how to leverage multiple data sources to build a telephony blacklist.
- Used topic modeling to learn the content of the unwanted calls.
- Evaluated the correctness and effectiveness of the blacklist.
- Our blacklist can block about 55% of the unwanted calls.
- Identified top running campaigns and analyzed their behaviour.
Thank You