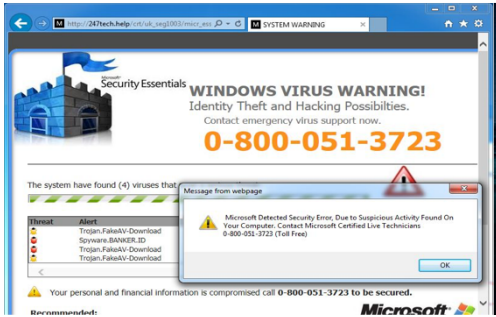
Towards Measuring the Effectiveness of Telephony Blacklists

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

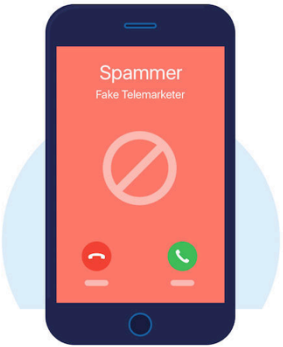
FTC

REC

honey pot

Blacklist Formation

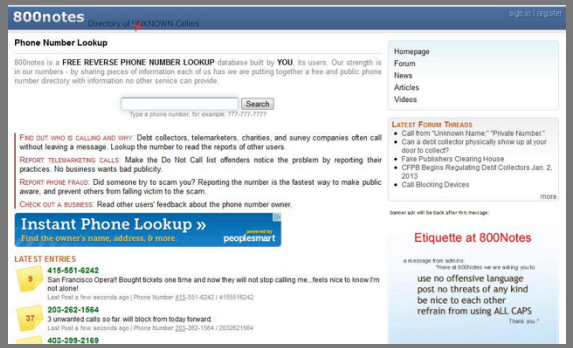
Evaluation



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.

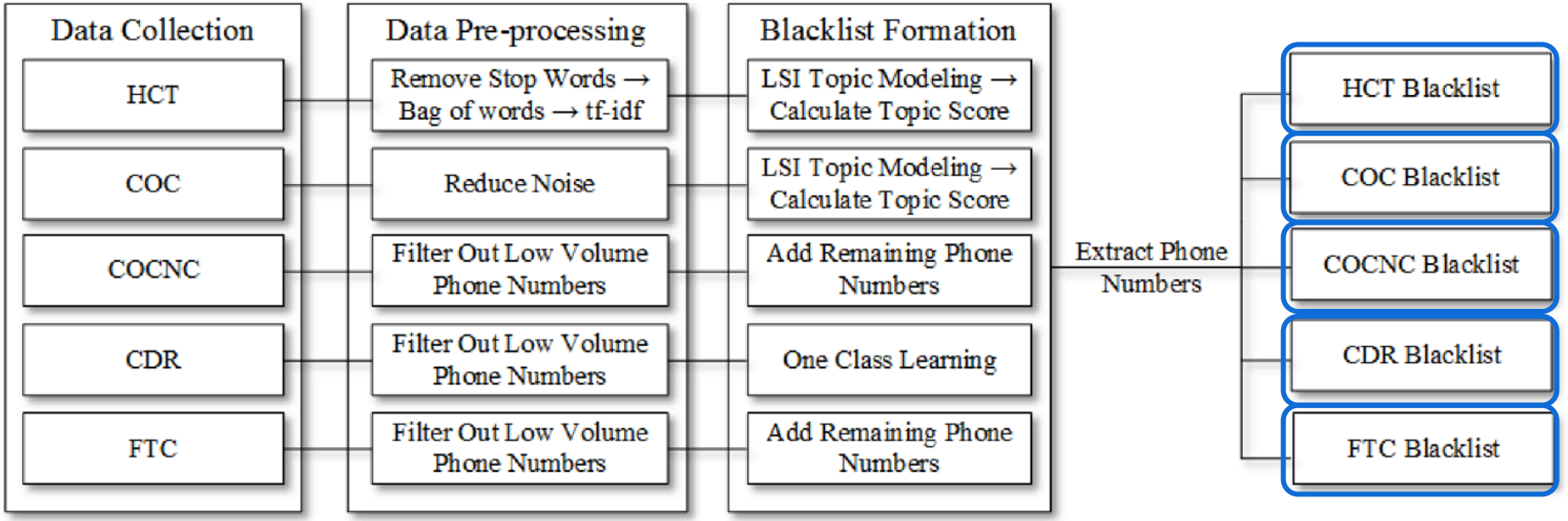
Dataset Summary

FTC		Number of FTC reports	1.56 million
		Number of Callers	300,000
CDR		Number of CDRs	1.1 million
		Number of Callers	200,000
		Number of Callees	58,000
COC		Number of User Comments	600,000
		Number of Callers	98,000
HCT		Number of Transcripts	20,000
		Number of Callers	9,500
		Number of Callees	6,000

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



Context-less Blacklisting

Blacklisting using the CDR data:

- Calculate blacklist score for each phone number p_i

$$s(p_i, \Delta t) = \alpha \times vol(p_i, \Delta t) + \beta \times nod(p_i, \Delta t)$$

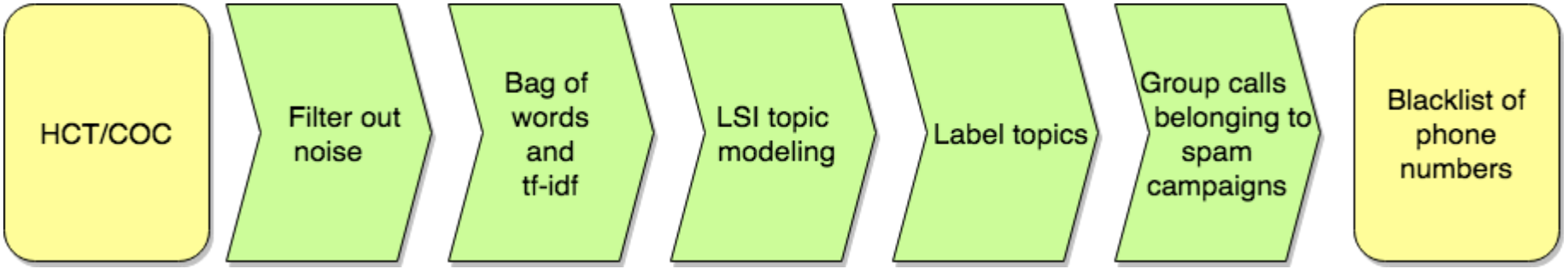
  Any number p_i whose blacklist score is greater than a predetermined threshold θ_b is added to the blacklist.

Blacklisting using the FTC/COCNC data:

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

Context-rich Blacklisting

Blacklisting using the HCT and COC data:

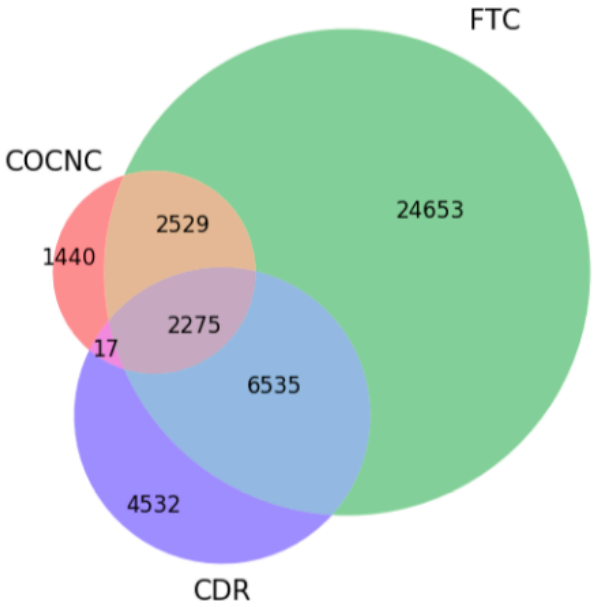


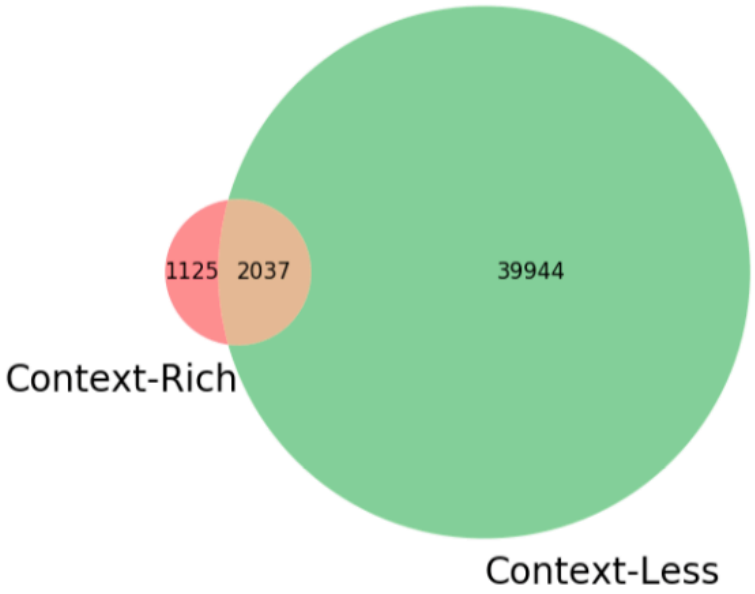
Topic modeling results

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

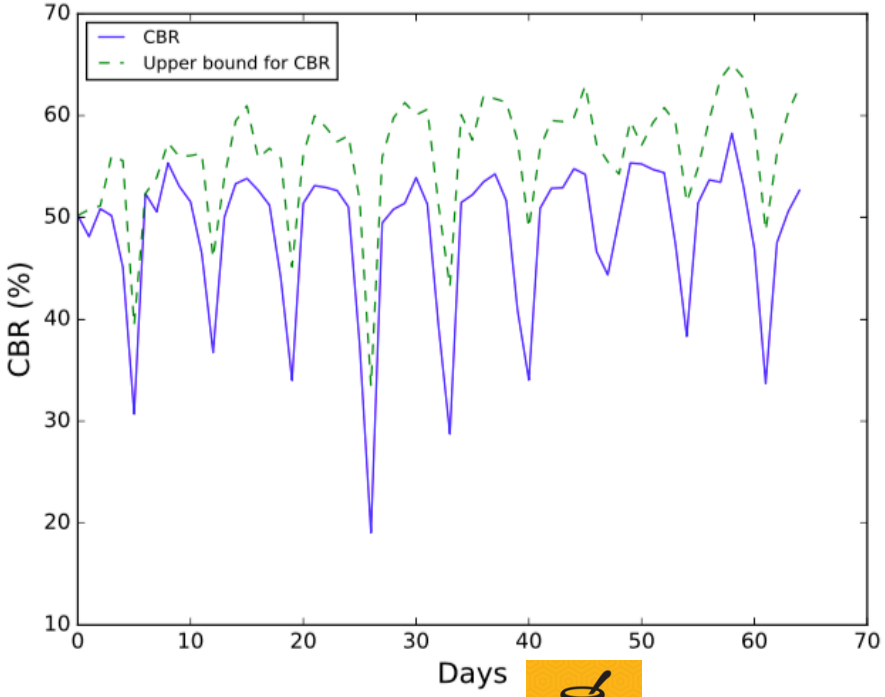
Results

Blacklist size and overlap



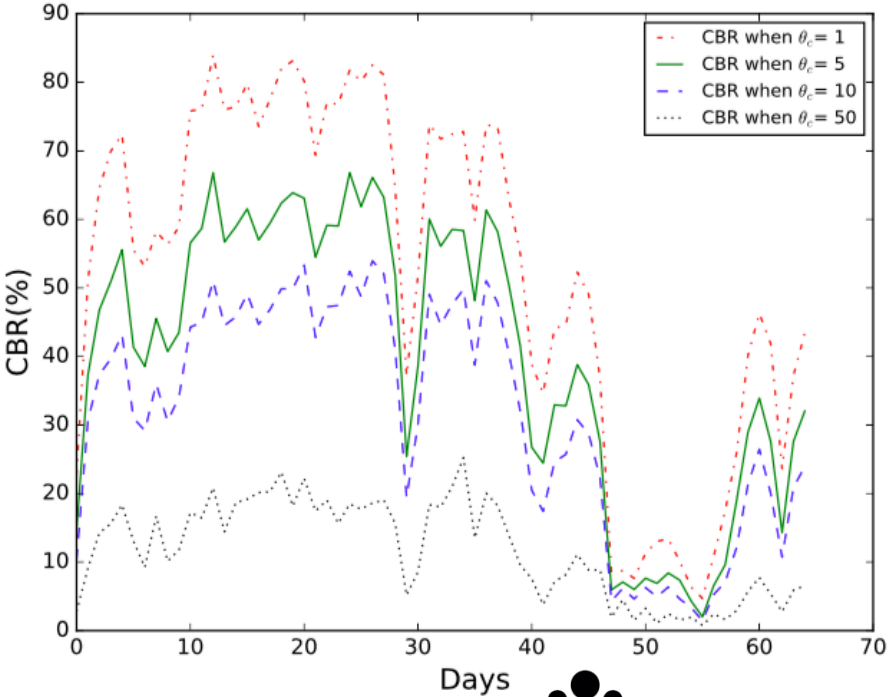


Call Blocking Rates



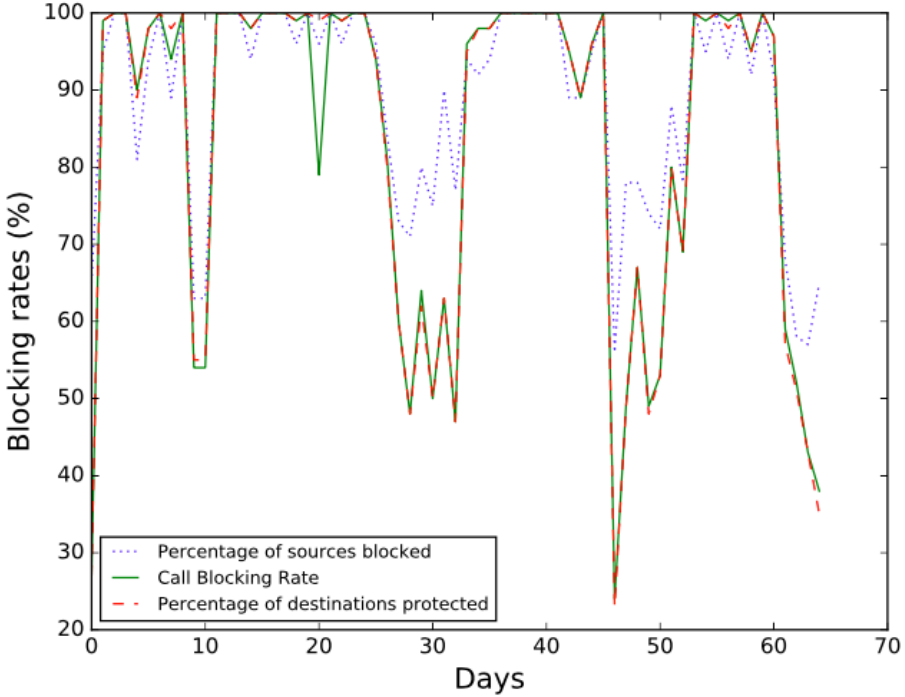
CDR

honey pot

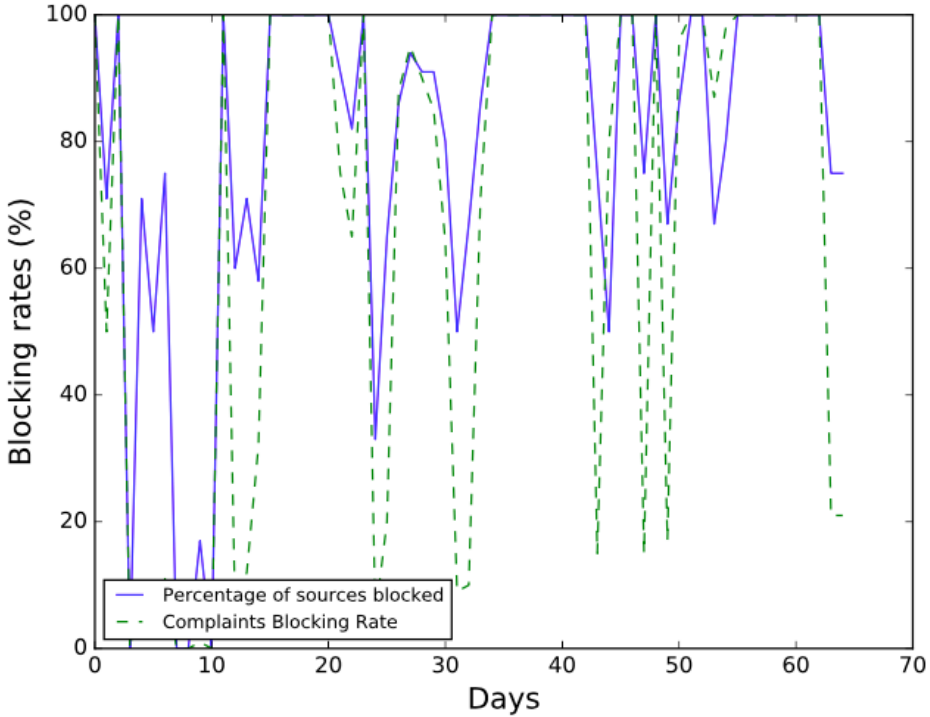


COCNC

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