Spam ain’t as Diverse as It Seems: Throttling OSN Spam with Templates Underneath

Hongyu Gao, Yi Yang, Kai Bu, Yan Chen, Doug Downey, Kathy Lee, Alok Choudhary

Northwestern University, USA

Zhejiang University, China

Background

Among world's most visited websites by Alexa

http://afrodigit.com/visited-websites-world/

2

1.35 billion monthly active users by Jul 2014

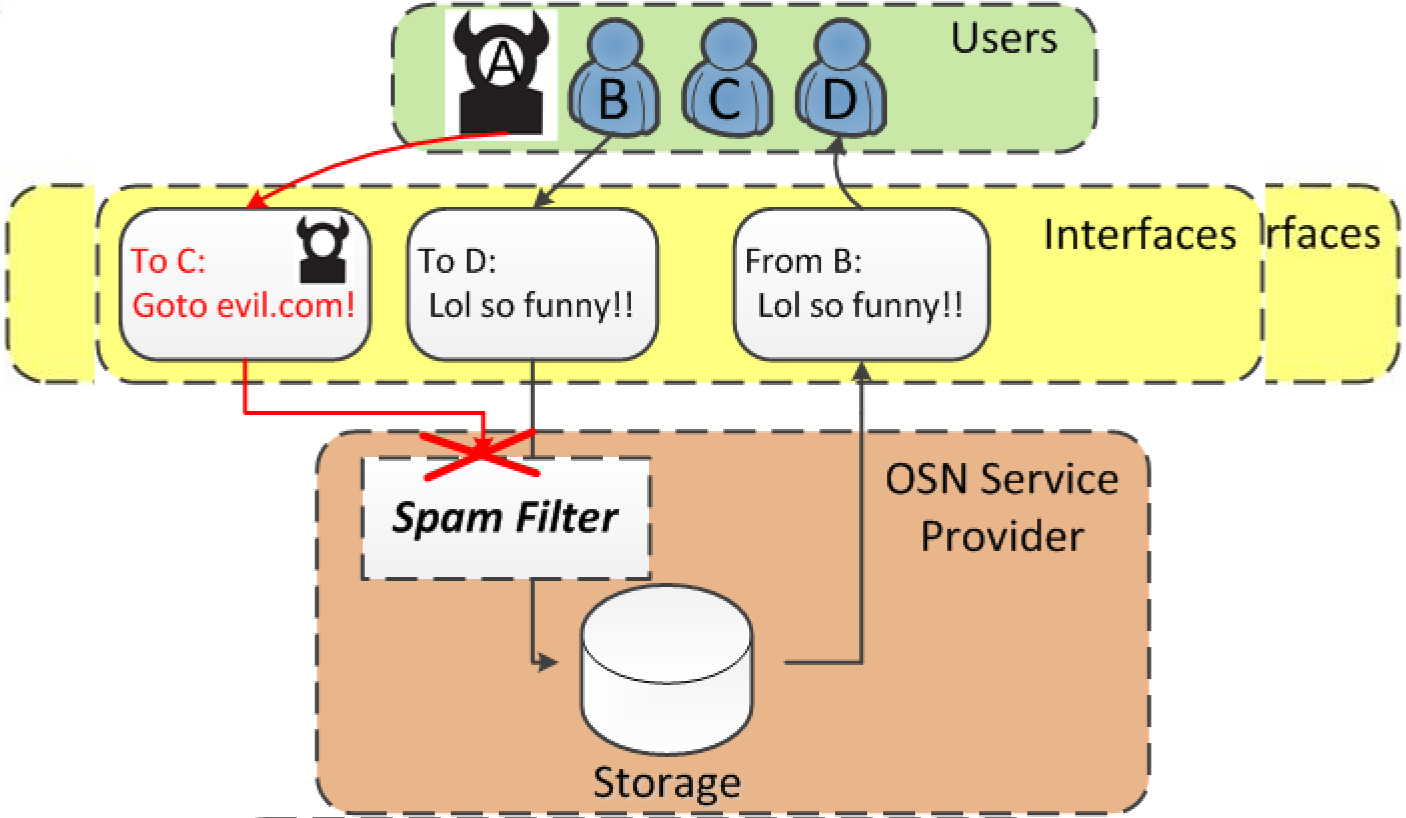
10

284 million users by Oct 2014

14

332 million users by Nov 2014

Background


Users
A
B
C
D
Interfaces
rfaces
To C:
Goto evil.com!
To D:
Lol so funny!!
From B:
Lol so funny!!
OSN Service
Provider
Spam Filter
Storage

Background

Scary Twitter spam stats

- 2011. **3.5 billion** tweets posted to Twitter every day are spam
  http://tinyurl.com/p8mqqvs
- 2014. **14 percent** of Twitter's user base is bots and spam bots
  http://tinyurl.com/l755bvm

Our Prior OSN Security Work

First study to offline detecting and characterizing Social Spam Campaigns (SIGCOMM IMC 2010)

- Largest scale experiment on Facebook then
  - 3.5M user profiles, 187M wall posts
- Confirm spam campaigns in the wild.
  - 200K spam wall posts in 19 significant campaigns.
- Featured in Wall Street Journal, MIT Technology Review and ACM Tech News

Online spam campaign discovery (NDSS 2012)

- Mostly use non-semantics information, syntactic clustering

How Are the Spam Tweets Generated?

Measuring Trend of Twitter Spam

- Download tweets containing popular hashtags
- Visit Twitter retrospectively to identify suspended accounts

- 2011 Twitter data:
  - 17 Million tweets
  - 558,706 spam tweets (>3%)

Template Model

- A macro sequence (m_1, m_2, …, m_k)
- Each macro instantiates differently during spam generation

Macro$_1$	**Macro$_2$**	**Macro$_3$**
Beppe Signori	making out with another man -	*URL*
Jason Isaacs	making out with another man -	*URL*
Beppe Signori	is really gay, look at this video	*URL*
Jason Isaacs	is really gay, look at this video	*URL*
RIP Jonas Bevacqua	is really gay, look at this video	*URL*

Template = *celebrity names* + *actions* + *URL*

Semi-automated Spam Measurement

Spam data	With Template	Paraphrase	No-content	Others
2011	63.0%	14.7%	8.4%	13.9%
2012	68.3%	12.9%	0.3%	18.5%

- The majority of spam is generated with underlying templates
- We collect a smaller 2012 Twitter data containing 46,891 spam tweets
- The prevalence of template-based spam is persistent

Syntactic only detection is not sufficient!

Semantics Based Spam Detection

- Extract spam template in real time
- Fight spam with its own template
- Detect multiple spam templates simultaneously

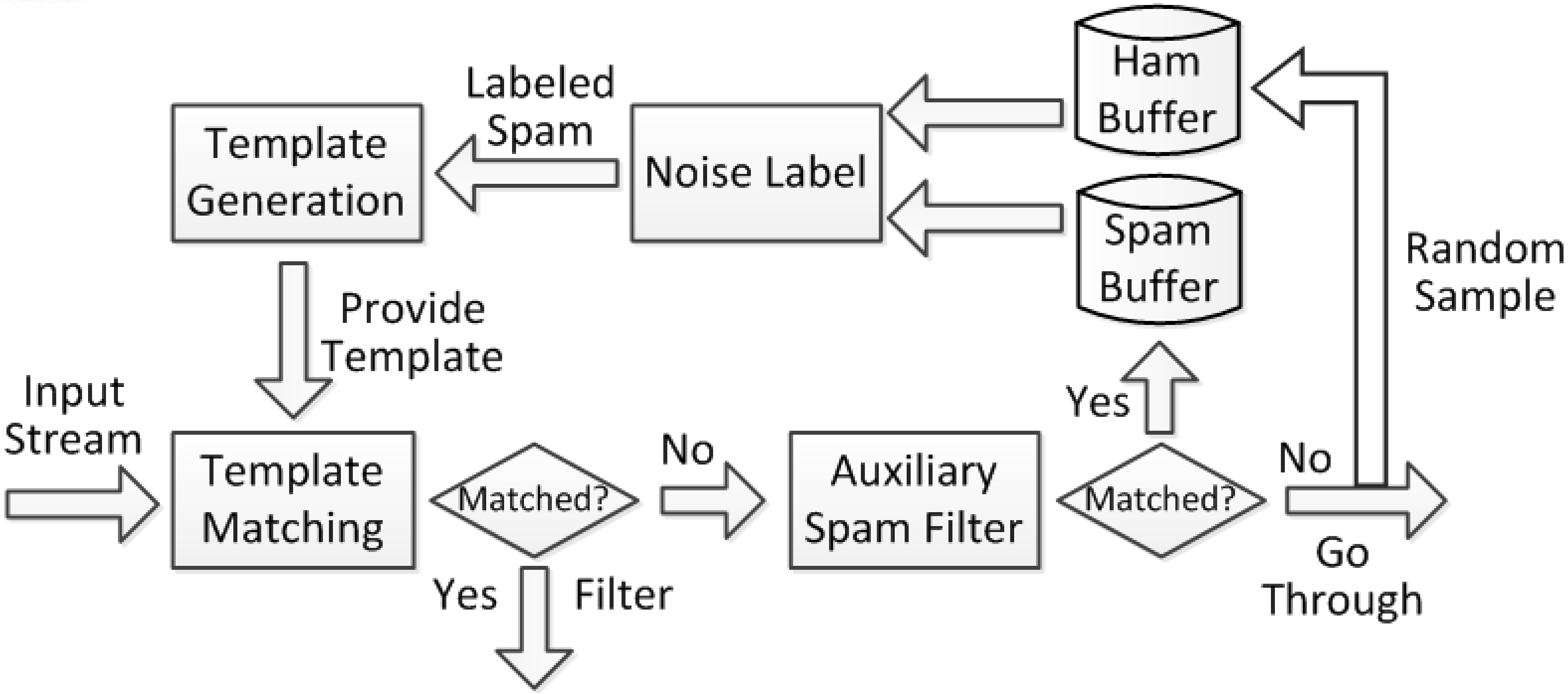
Challenges

- Absence of invariant substring in template
  - Prior study assumes the existence of invariant substrings. [Pitsillidis NDSS'10][Zhang NDSS'14]
- Prevalence of noise
  - Spammers extensively add semantically unrelated noise words into spam messages.
- Spam heterogeneity
  - It is hard to obtain a training set containing spam instantiating a single template in practice.

Solutions

- Absence of invariant substring in template
  – Spam template generation without the need for invariant substring.
- Prevalence of noise
  – Automated noise labeling to identify and exclude noise words from template generation.
- Spam heterogeneity
  – Cluster and refine.

Template Generation/Matching Module



- Real-time detection
- The auxiliary spam filter supplies training spam samples
  - Could use black list or any other spam detection systems
  - Heterogeneous filters to avoid evasion

Single Campaign Template Generation

Step 1: Compute a “good” common super-sequence

(Majority-Merge algorithm)

Beppe Signori making out – URL
Jason Isaacs making out – URL
Beppe Signori is really gay URL
Jason Isaacs is really gay URL
RIP Jonas Bevacqua is really gay URL

Super-sequence

Beppe	**Signori**	**Jason**	**Isaacs**	**making**	**out**	**is**	**really**	**gay**	**-**	**url**	**RIP**	**Jonas**	**Bevacqua**	**is**	**really**	**gay**	**url**
Beppe	Signori	ε	ε	making	out	ε	ε	ε	-	url	ε	ε	ε	ε	ε	ε	ε
ε	ε	Jason	Isaacs	making	out	ε	ε	ε	-	url	ε	ε	ε	ε	ε	ε	ε
Beppe	Signori	ε	ε	ε	ε	is	really	gay		ε	ε	ε	ε	ε	ε	ε	url
ε	ε	Jason	Issacs	ε	ε	is	really	gay		ε	ε	ε	ε	ε	ε	ε	url
ε	ε	ε	ε	ε	ε	ε	ε	ε		ε	RIP	Jonas	Bevacqua	is	really	gay	url

Single Campaign Template Generation

Step 2: Matrix columns reduction

Beppe	Signori	Jason	Isaacs	making	out	is	really	gay	-	url	RIP	Jonas	Bevacqua	is	really	gay	url
Beppe	Signori	ε	ε	making	out	ε	ε	ε	-	url	ε	ε	ε	ε	ε	ε	ε
ε	ε	Jason	Isaacs	making	out	ε	ε	ε	-	url	ε	ε	ε	ε	ε	ε	ε
Beppe	Signori	ε	ε	ε	ε	is	really	gay		ε	ε	ε	ε	ε	ε	ε	url
ε	ε	Jason	Issacs	ε	ε	is	really	gay		ε	ε	ε	ε	ε	ε	ε	url
ε	ε	ε	ε	ε	ε	ε	ε	ε		ε	RIP	Jonas	Bevacqua	is	really	gay	url

Super-sequence

(Beppe| ε) (Signori| ε) (Jason| ε) (Isaacs| ε) ...

Beppe	Signori	Jason	Isaacs	making	out	-	RIP	Jonas	Bevacqua	is	really	gay	url
Beppe	Signori	ε	ε	making	out	-	ε	ε	ε	ε	ε	ε	url
ε	ε	Jason	Isaacs	making	out	-	ε	ε	ε	ε	ε	ε	url
Beppe	Signori	ε	ε	ε	ε	ε	ε	ε	ε	is	really	gay	url
ε	ε	Jason	Issacs	ε	ε	ε	ε	ε	ε	is	really	gay	url
ε	ε	ε	ε	ε	ε	ε	RIP	Jonas	Bevacqua	is	really	gay	url

Single Campaign Template Generation

Step 3: Matrix columns concatenation

Beppe	Signori	Jason	Isaacs	making	out	-	RIP	Jonas	Bevacqua	is	really	gay	url
Beppe	Signori	ε	ε	making	out	-	ε	ε	ε	ε	ε	ε	url
ε	ε	Jason	Isaacs	making	out	-	ε	ε	ε	ε	ε	ε	url
Beppe	Signori	ε	ε	ε	ε	ε	ε	ε	ε	is	really	gay	url
ε	ε	Jason	Issacs	ε	ε	ε	ε	ε	ε	is	really	gay	url
ε	ε	ε	ε	ε	ε	ε	RIP	Jonas	Bevacqua	is	really	gay	url

Regular Expression Template

Beppe Signori|Jason Isaacs|RIP Jonas Bevacqua	Is really gay|making out -	url
Beppe Signori	making out -	url
Jason Isaacs	making out -	url
Beppe Signori	is really gay	url
Jason Issacs	is really gay	url
RIP Jonas Bevacqua	is really gay	url

Solutions

- Spam template generation without the need for invariant substring.

- Automated noise labeling to identify and exclude noise words from template generation.

- Cluster and refine for mixture of spam campaigns.

Noise Labeling

Key problem: spammers extensively insert noise words into spam messages

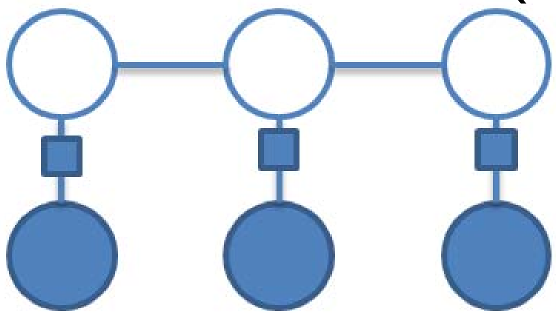
- To draw a larger audience
- To diversify the message

@mentions, *#hashtags*, *popular terms*, *etc.*

Noise Labeling

Goal: exclude the noise words from the template generation process.

Method: treat noise detection as a sequence labeling task, using Conditional Random Fields (CRFs) approach.



Output: a “noise” or “non-noise” label for each word in the message.

Feature Selection

Intuition: noise words are popular, but the combination of them are not popular.

Features:

- $freq(t_i)$
- $freq(t_i t_{i+1})^2/(freq(t_i)freq(t_{i+1}))$
- $freq(t_{i-1} t_i)^2/(freq(t_{i-1})freq(t_i))$

Orthographic features:

- *Is capitalized?*
- *Is hashtag?*
- *Is numeric?*
- *Is user mention?*

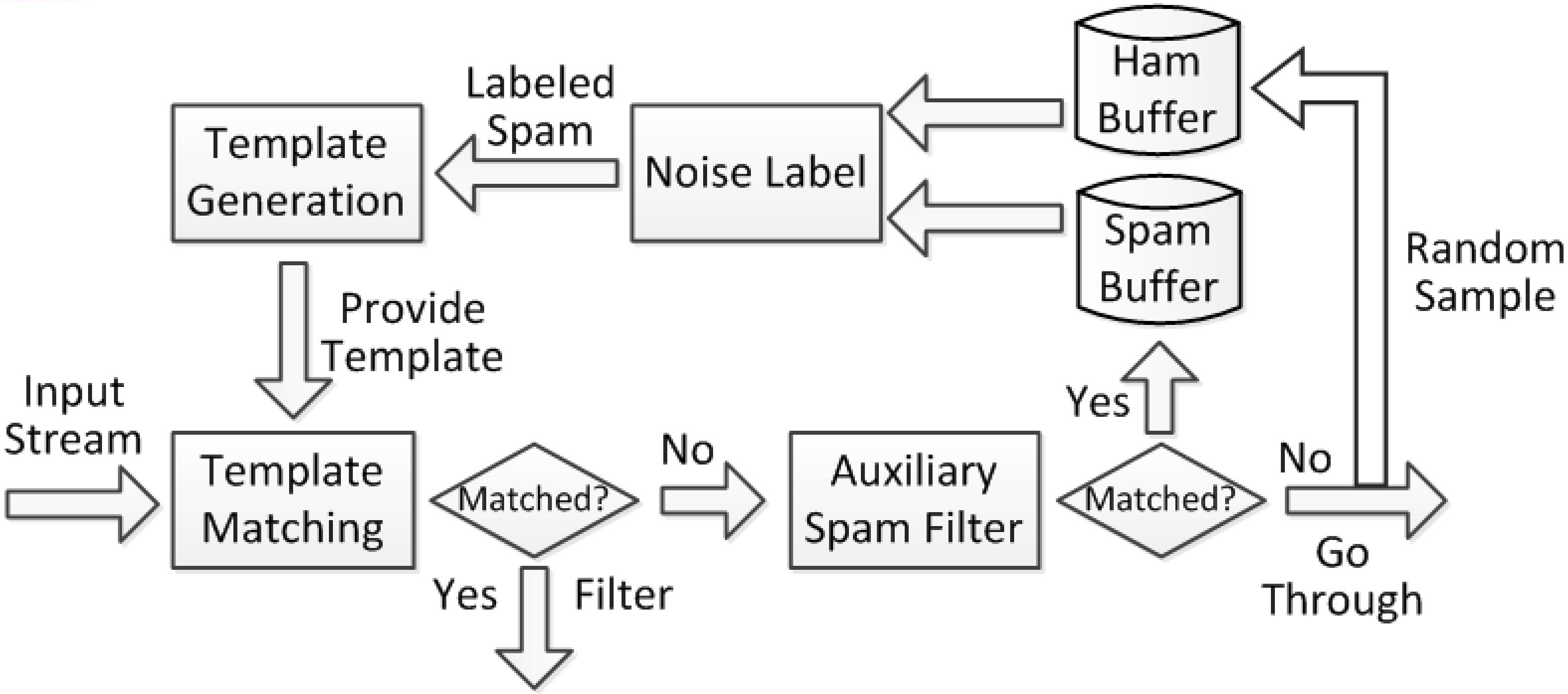
Solutions

- Spam template generation without the need for invariant substring.
- Automated noise labeling to identify and exclude noise words from template generation.
- **Cluster and refine for mixture of spam campaigns.**

Multi-campaign Template Generation

- Problem: in realistic scenario the system observes the mixture of spam instantiating multiple templates, rather than a single one.
- Solution:
  - Part 1, coarse pre-clustering, using standard clustering technique.
  - Part 2, refine the single campaign template generation process, by limiting the ratio of “ ε ” in the matrix to prune out “outlier” messages.

Recap: Template Generation/Matching Module



- Real-time detection
- The auxiliary spam filter supplies training spam samples

Evaluation Results

- Dataset:
  - 17M tweets generated between June 1, 2011 and July 21, 2011
  - 558,706 spam tweets

- Auxiliary spam filter:
  - The online campaign discovery module (introduced later)
  - 63.3% TP rate, 0.27% FP rate

Detection Accuracy

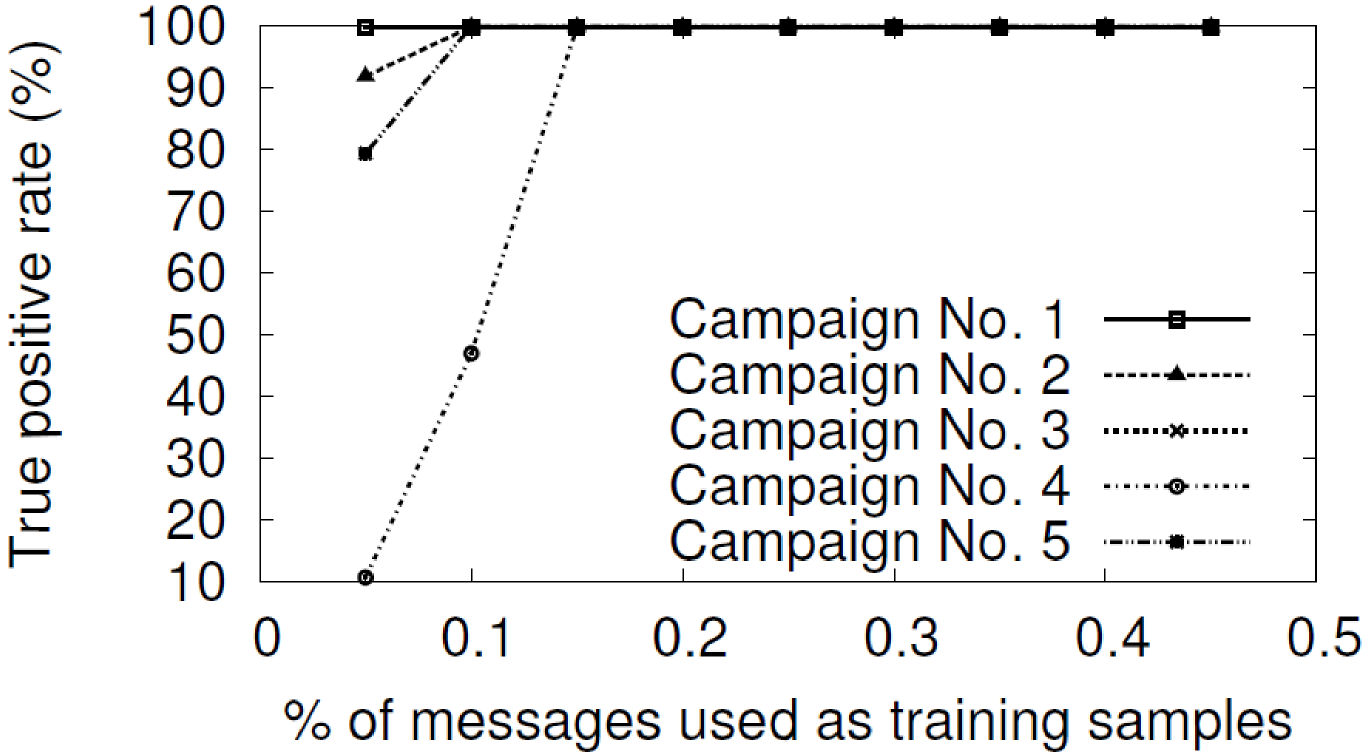
Module	Template Generation	Auxiliary Filter	Combined
Spam Category			
Template-based	95.7%	70.1%	98.4%
Paraphrase	51.0%	51.4%	70.1%
No-content	73.8%	67.0%	83.1%
Others	18.4%	43.2%	44.7%
Overall TP	76.2%	63.3%	85.4%
FP	0.12%	0.27%	0.33%

Generated Template Example

Top 5 generated templates with the most matching spam:

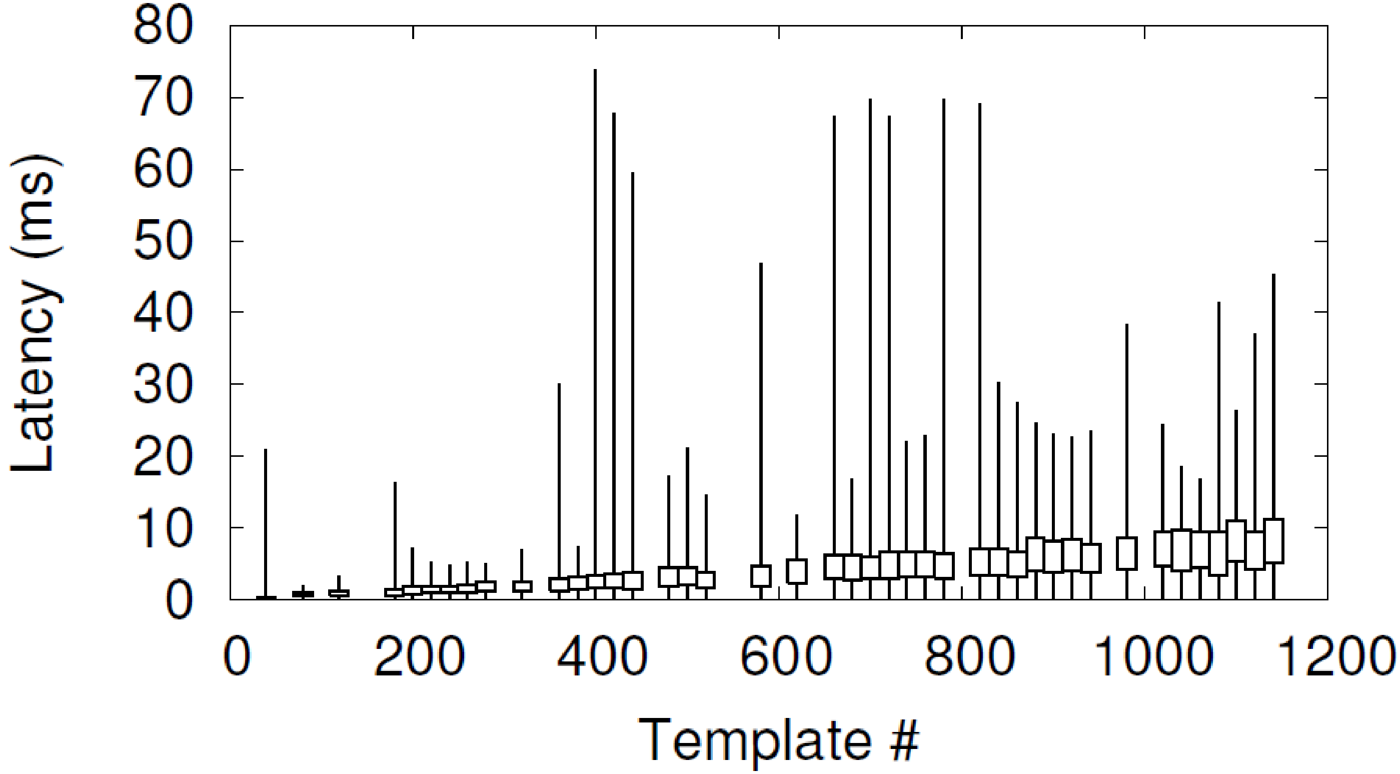
Spam #	Template
11.1%	^ (I wager|My my ,) you (cannot| ε) (ε |defeat) this \\. URL .* $
7.2%	^ The (ε |folks|people) at my (ε |place|location) are groveling for this ! URL .* $
6.4%	^ You (will not|won't| ε) (ε |think|believe) this \\. The (ε |best|greatest) (thing|factor| ε) (because|since) slice bread \\.
5.0%	^ (Cool|Wow|Amazing) , I (by no means|in no way) (found|noticed) (people|anyone) (do that| ε) (just before|prior to) \\. URL .* $
4.1%	^ You (will not|won't| ε) (think|believe| ε) the (issues|points|things) they do on this (site|web page|web-site) \\. URL .* $

Sensitivity for New Campaigns



- Pick the top 5 campaigns
- All campaigns achieve almost 100% detection rate with 0.15% of messages as training samples.
- The system can react to newly emerged campaigns quickly.

Template Matching Speed



- The median matching latency grows slowly with template number, less than 8ms.
- The largest latency is less than 80ms, unnoticeable to users.

Conclusions

- Tangram: first system to real time extract multiple spam templates without unique invariants.
  - 63% of Twitter spam is generated by templates.
  - Detect 95.7% of template-based spam.
  - Overall TP rate of 85.4% and FP rate of 0.33%.
- Applying text analytics in other security applications
  - Measuring the Description-to-permission Fidelity in Android Applications, CCS 2014

Existing Work, cont'd

- Spam template generation [Pitsillidis NDSS'10][Zhang NDSS'14]
  - How to detect spam without invariant substrings?

- Spammer account detection [Stringhihi ACSAC'10][Yang RAID'11]
  - How to detect spam in real-time?
  - How to detect spam originating from compromised accounts, e.g., in a worm propagation scenario?

Thank you!

http://list.cs.northwestern.edu/

Questions?

Background

Filtering Twitter spam is uniquely challenging

- Twitter exposes developer APIs to make it easy to interact with Twitter platform
- Real-time content is fundamental to Twitter user's experience

http://tinyurl.com/oxtmmnz