Towards Online Spam Filtering in Social Networks

Hongyu Gao, Yan Chen, Kathy Lee, Diana Palsetia and Alok Choudhary

Lab for Internet and Security Technology (LIST)
Department of EECS
Northwestern University

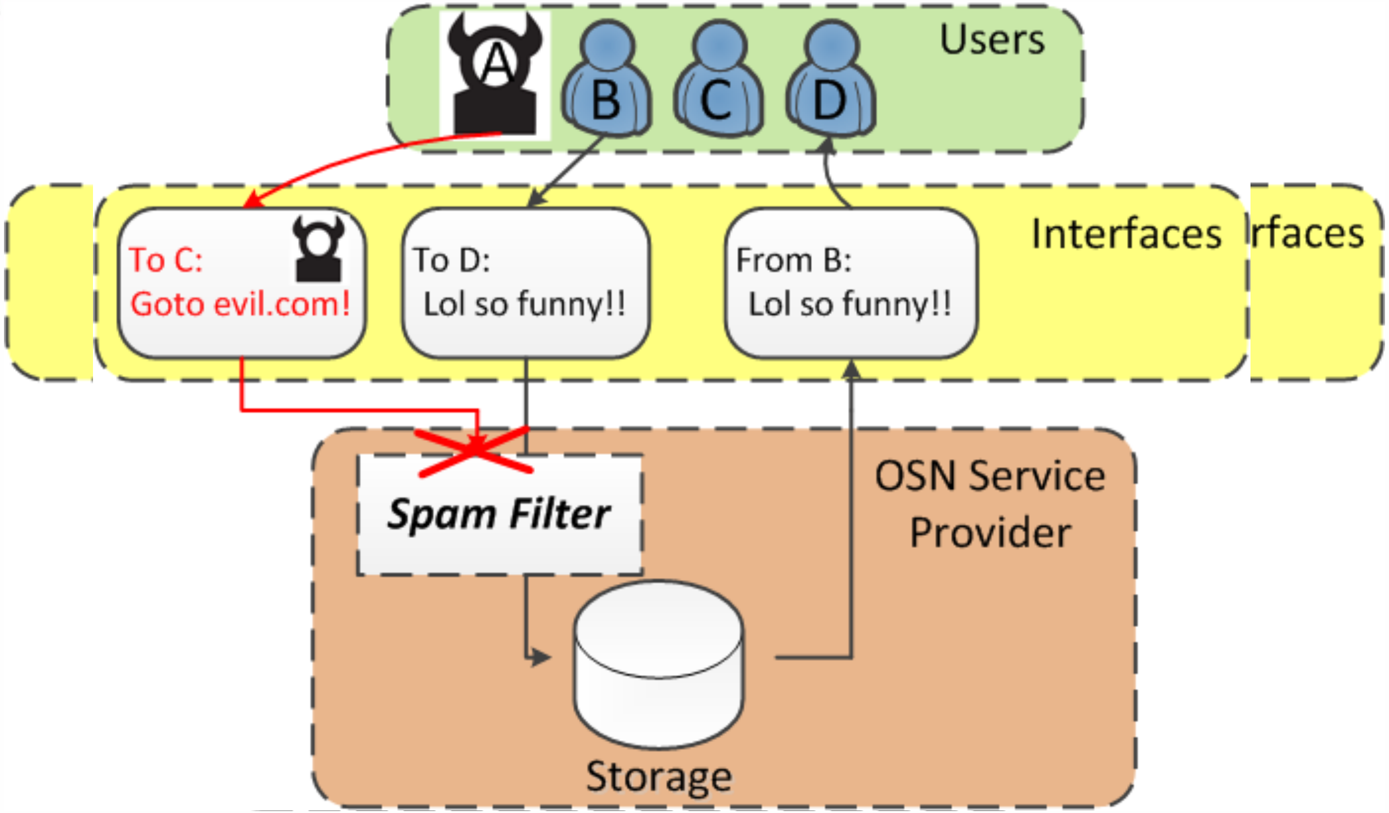
Background

People on Facebook	More than 800 million active users
	More than 50% of our active users

2 **Facebook**
facebook.com
A social utility that connects people, to keep up with friends, upload photos, share links and ... More
★★★★☆ Search Analytics ▶ Audience ▶

9 **Twitter**
twitter.com
Social networking and microblogging service utilising instant messaging, SMS or a web interface.
★★★★☆ Search Analytics ▶ Audience ▶

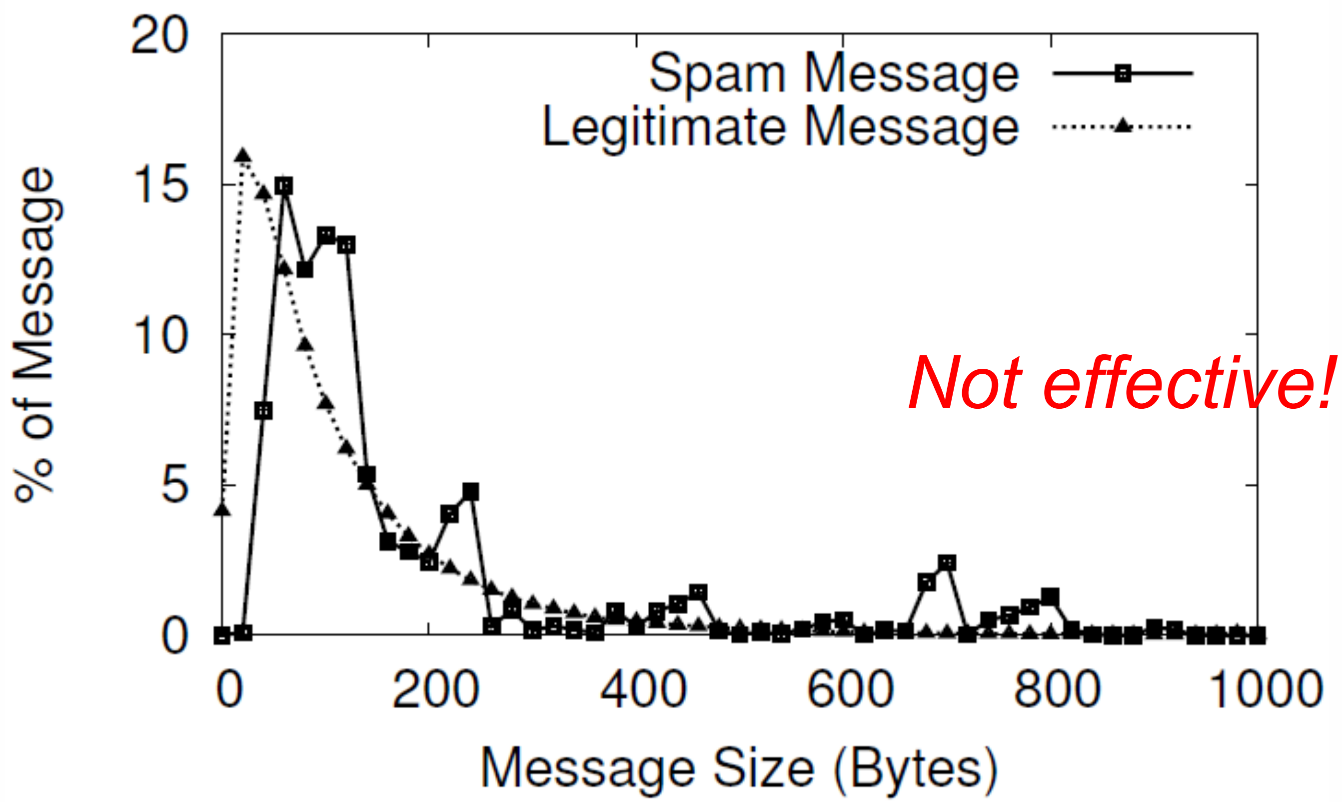
Background



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

Another Study in Spam Detection??

- Unique characteristics of OSNs
  - Are existing features still effective?



Goals and Existing Work

- An effort towards a system ready to deploy
  - ❖ Online detection
  - ❖ High accuracy
  - ❖ Low latency
  - ❖ Detection of campaigns absent from training set
  - ❖ No need for frequent re-training
- Existing studies in OSN spam:
  - [Gao IMC10, Grier CCS10] offline analysis
  - [Thomas Oakland11] landing page *vs.* message content
  - Numerous work in spammer-faked account detection

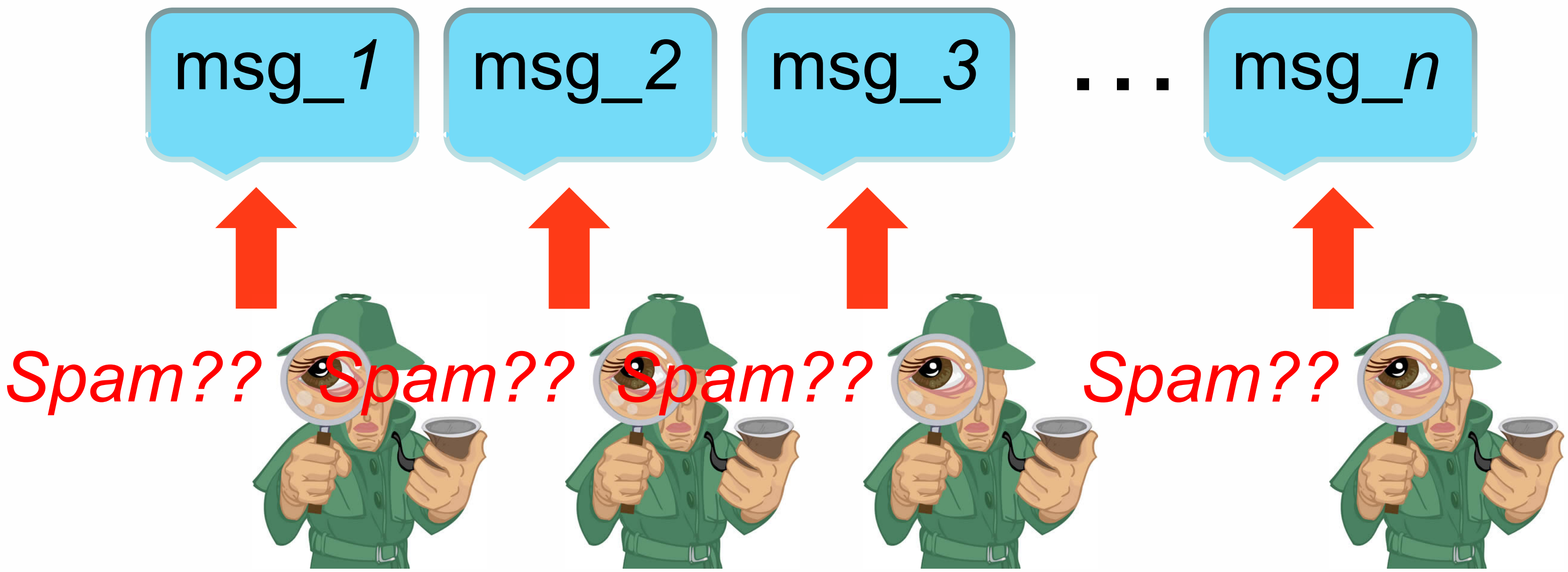
Roadmap

- **Detection System Design**
- Evaluation
- Conclusions & Future Work

Key Intuition

We Do NOT:

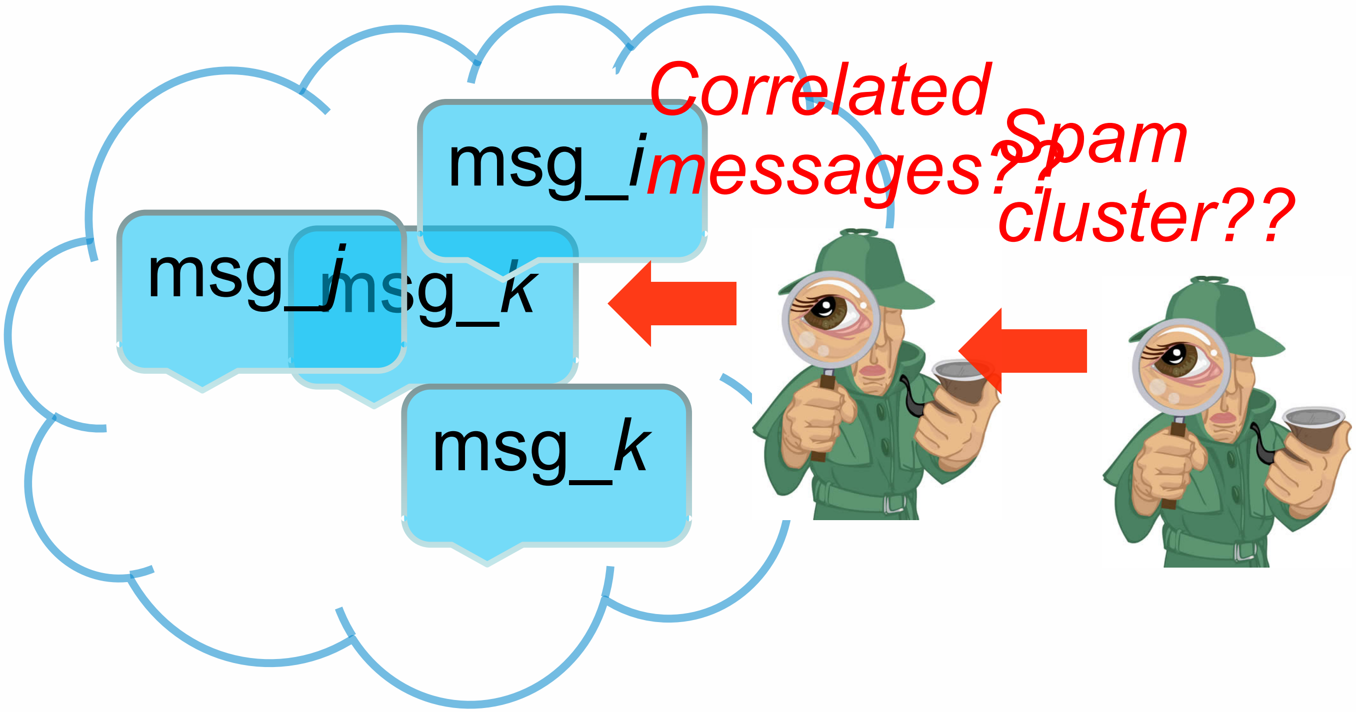
Inspect each message individually



Key Intuition

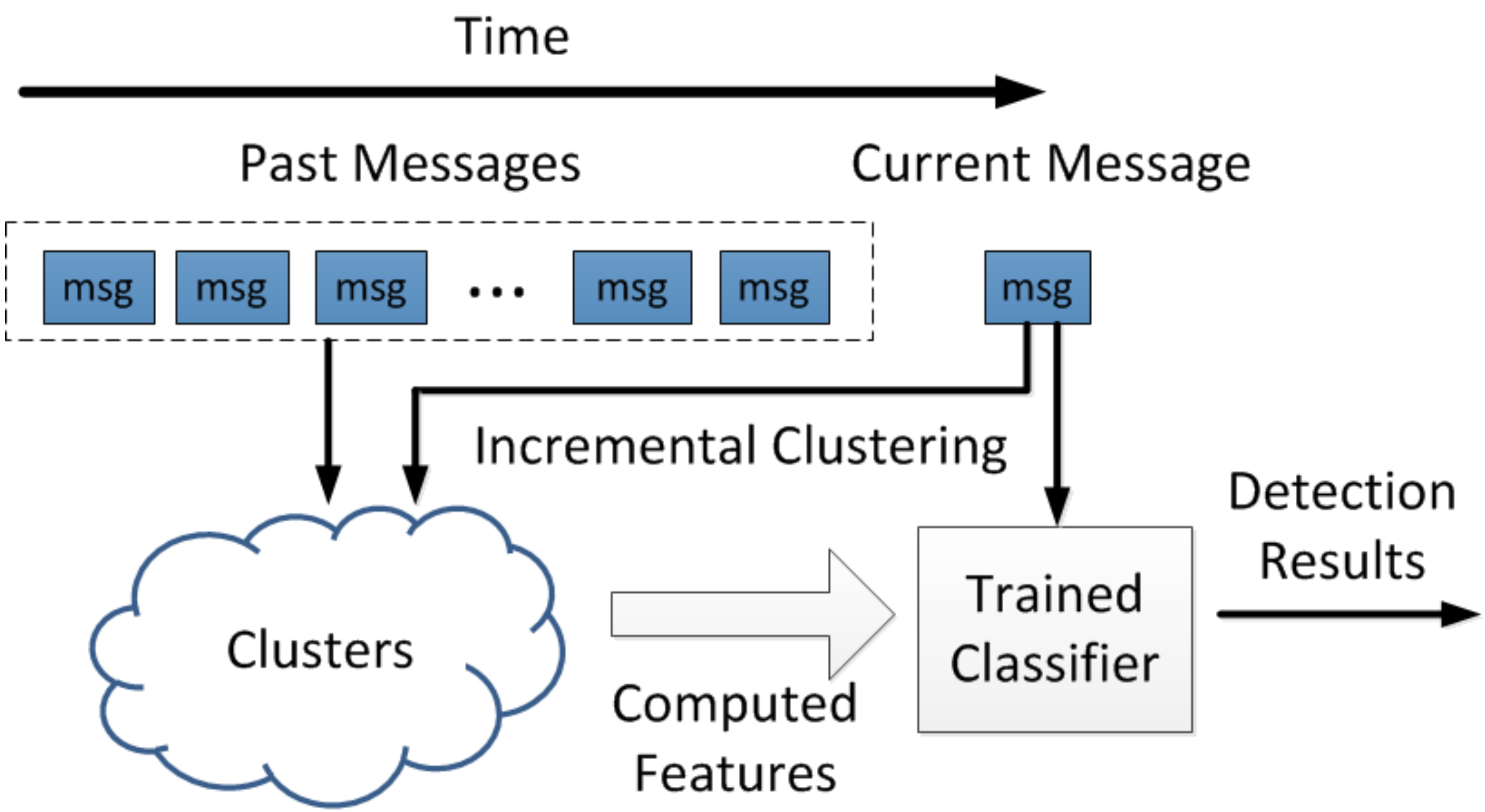
We Do:

Inspect correlated message clusters



System Overview

Detect coordinated spam campaigns.



Incremental Clustering

- Requirement:
  - Given the clustering result of the first k messages and $(k+1)_{th}$ message
  - Efficiently compute the result of the (k+1) messages
- Adopt text shingling technique
  - Pros: High efficiency
  - Cons: Syntactic method

Feature Selection

- Feature selection criteria:
  - Cannot be easily maneuvered.
  - Grasp the commonality among campaigns.
- 6 identified features:

❖ Sender social degree
❖ Interaction history
❖ Cluster size
❖ Average time interval
❖ Average URL #
❖ Unique URL #

Roadmap

- Detection System Design
- **Evaluation**
- Conclusions & Future Work

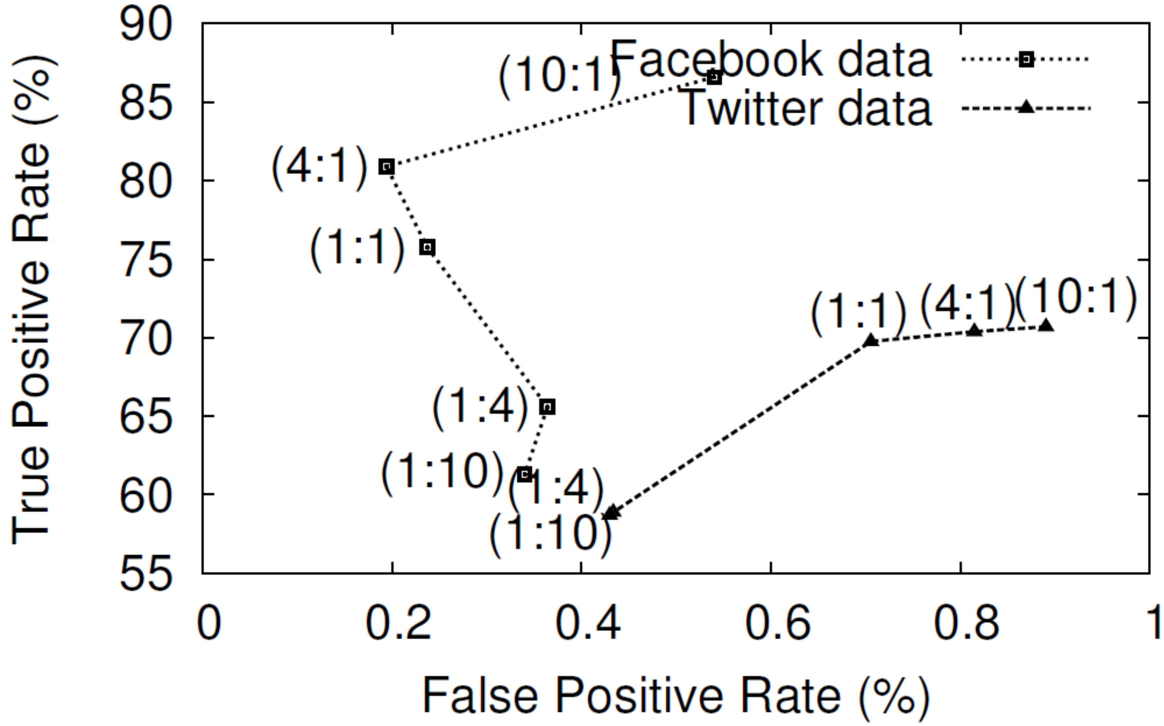
Dataset and Method

Site	Size	Spam #	Time
Facebook	187M	217K	Jan. 2008 ~ Jun. 2009
Twitter	17 M	467K	Jun. 2011 ~ Jul. 2011

- All experiments obey the time order
  - First 25% as training set, last 75% as testing set.
- Evaluated metrics:

❖ Overall accuracy
❖ Accuracy of feature subset
❖ Accuracy over time
❖ Accuracy under attack
❖ Latency
❖ Throughput

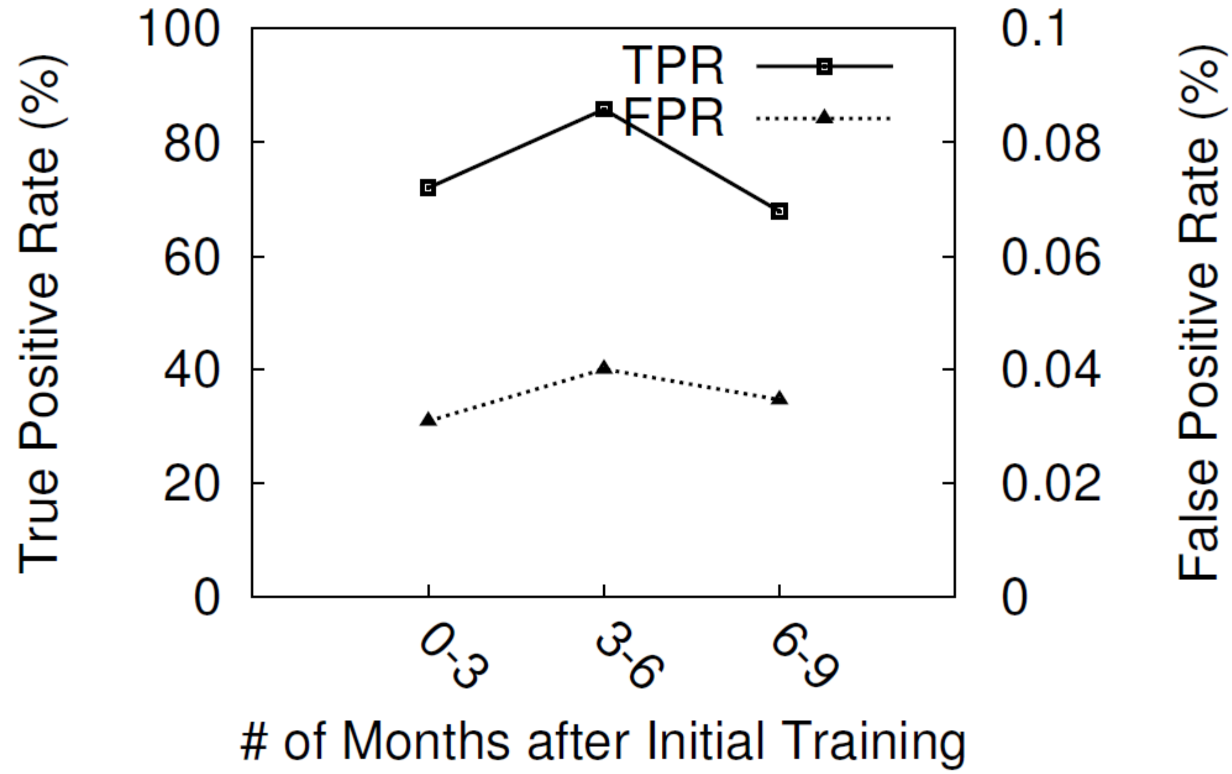
Overall Accuracy



Best result

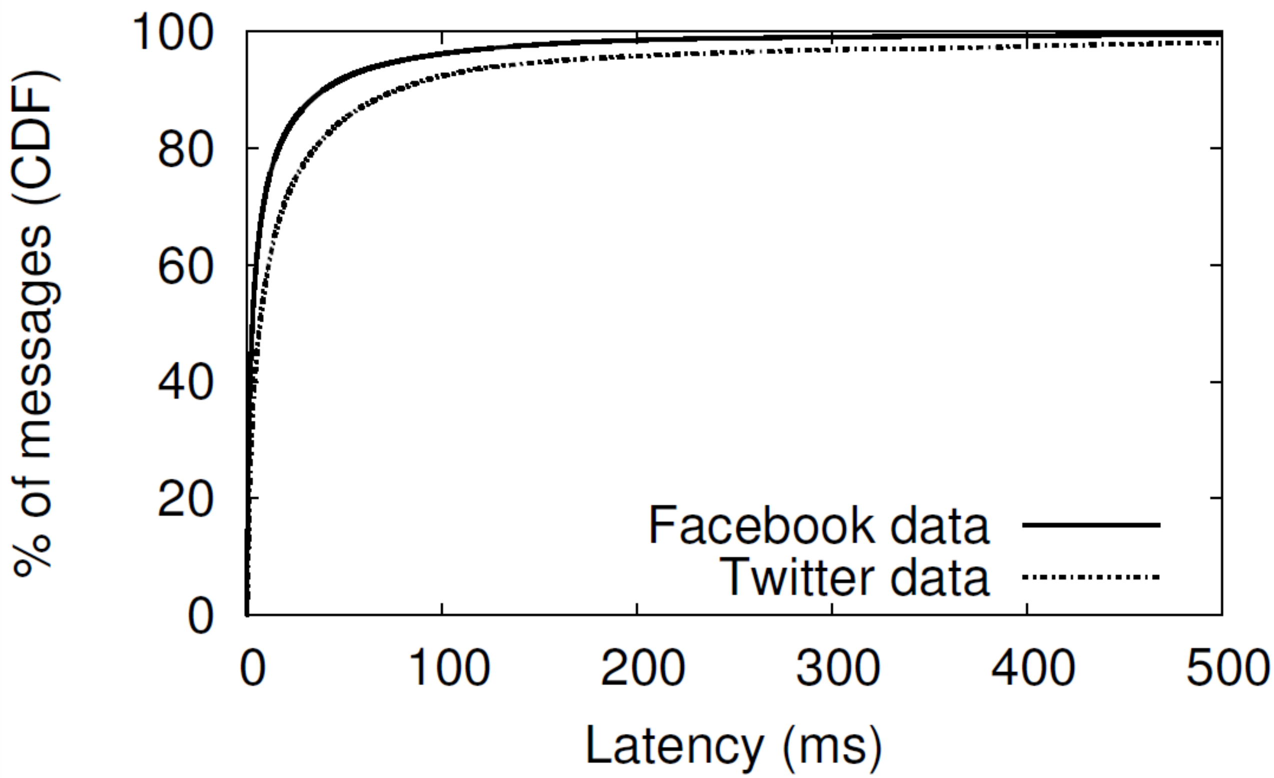
- FB: 80.9% TP 0.19%FP
- TW: 69.8%TP 0.70%FP

Accuracy over Time



No significant drop of TP or increase of FP

Latency



Latency (ms)	Facebook	Twitter
Mean	21.5	42.6
Median	3.1	7.0

Roadmap

- Detection System Design
- Evaluation
- **Conclusions &** Future Work

Conclusions

- We design an online spam filtering system based on spam campaigns.
  - Syntactical incremental clustering to identify message clusters
  - Supervised machine learning to classify message clusters
- We evaluate the system on both Facebook and Twitter data
  - 187M wall posts, 17M tweets
  - 80.9% TPR, 0.19% FPR, 21.5ms mean latency

Prototype release:

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

Future Work

Cool	, I	by no means	noticed	anyone	do that	prior to	. {URL}
Wow	, I	in no way	noticed	anyone		just before	. {URL}
Amazing	, I	by no means	found	people	do that	just before	. {URL}

Call for semantic clustering approaches

{Cool | Wow | Amazing} + , I + {by no means | in no way} + {noticed | found} + {anyone | people} + {do that | ε} + {prior to | just before} + . {URL}

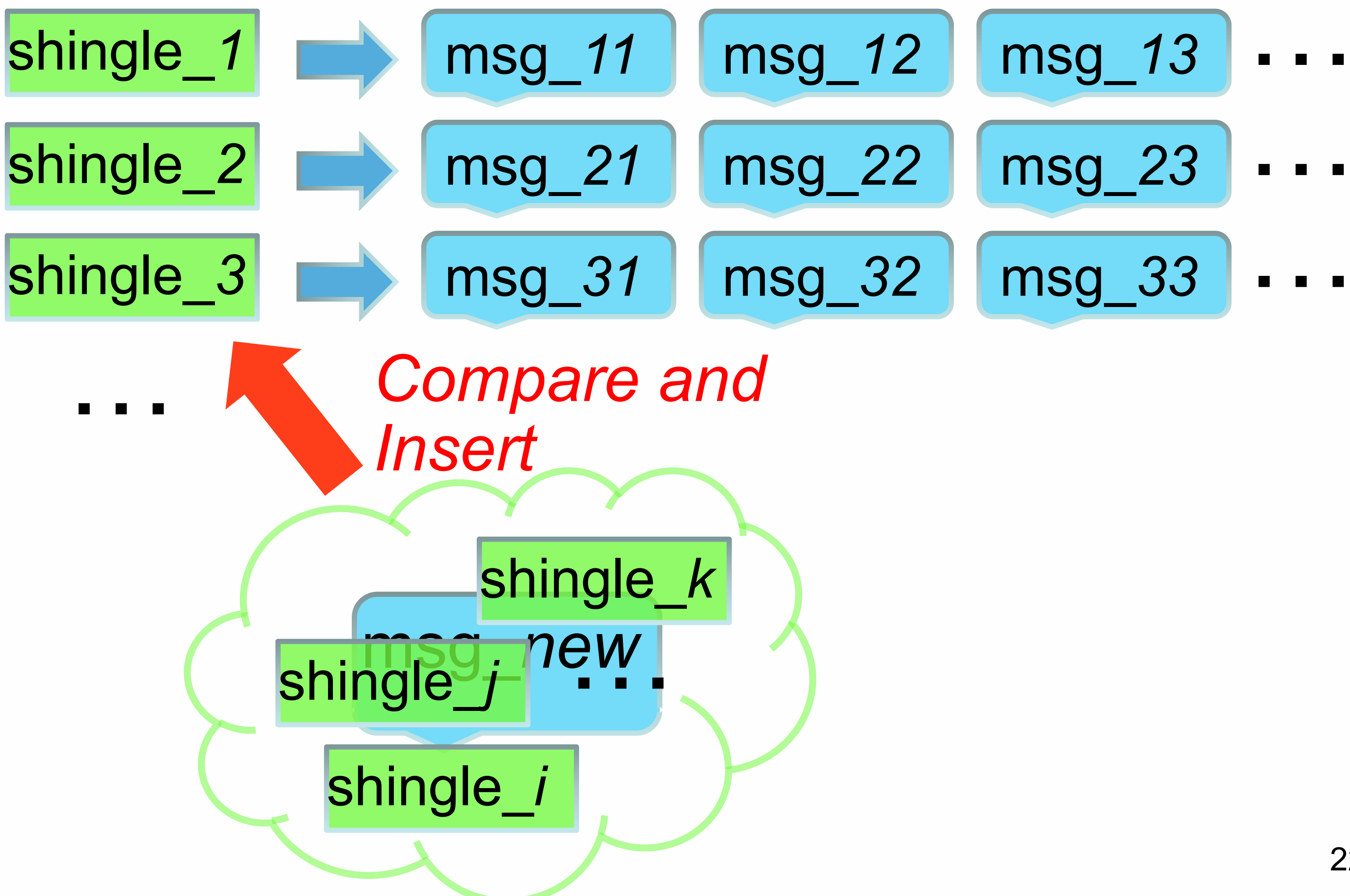
Template generation?

Thank you!

Contributions

- Design an online spam filtering system to deploy as a component of the OSN platform.
  - High accuracy
  - Low latency
  - Tolerance for incomplete training data
  - No need for frequent re-training
- Release the system
  - http://list.cs.northwestern.edu/socialnetworksecurity

Incremental Clustering

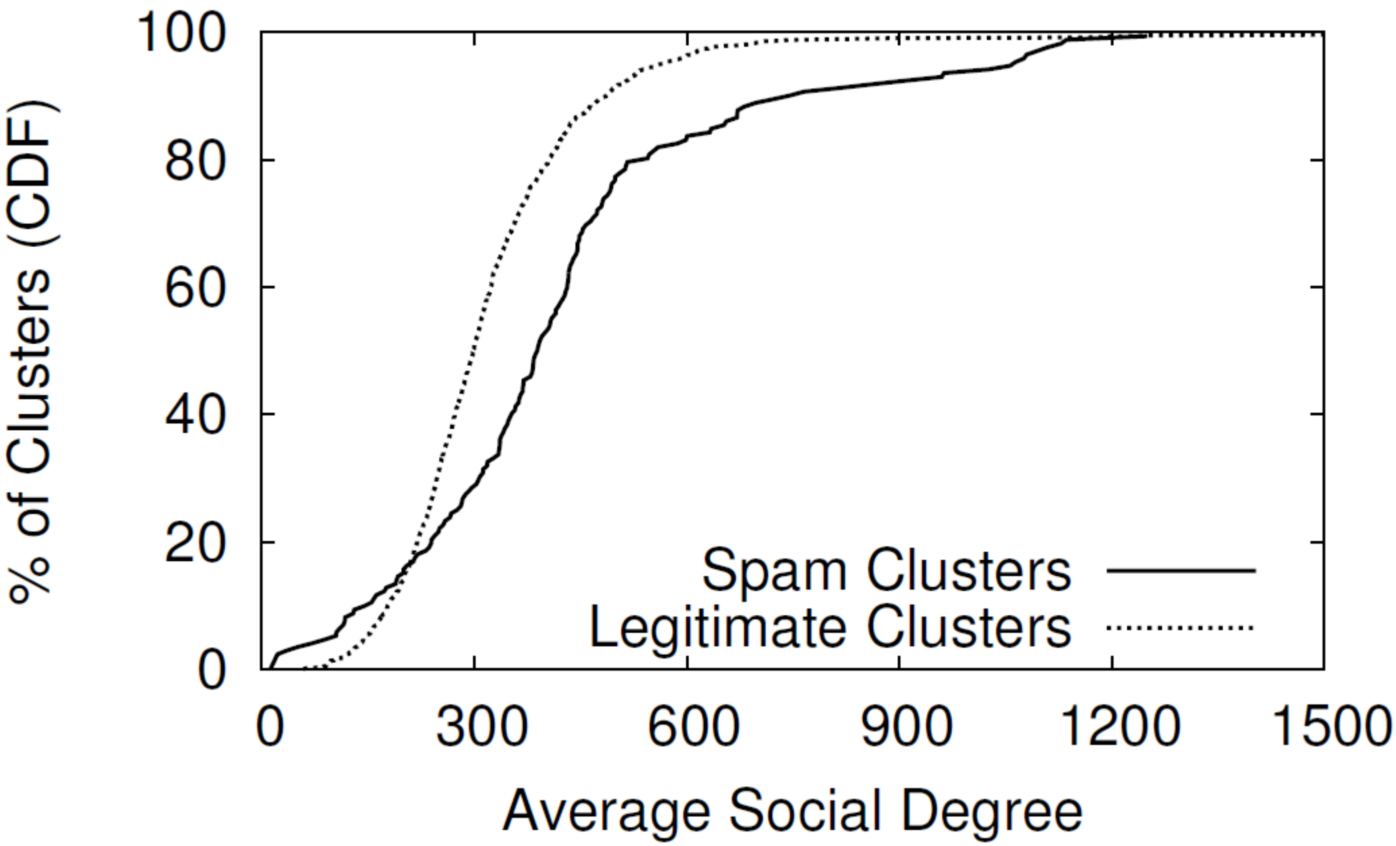


Sender Social Degree

- Compromised accounts:
  - The more edges, with a higher probability the node will be infected quickly by an epidemic.
- Spammer accounts:
  - Social degree limits communication channels.
- Hypothesis:
  - Senders of spam clusters have higher average social degree than those of legitimate message clusters.

Sender Social Degree

Average social degree of spam and legitimate clusters, respectively.

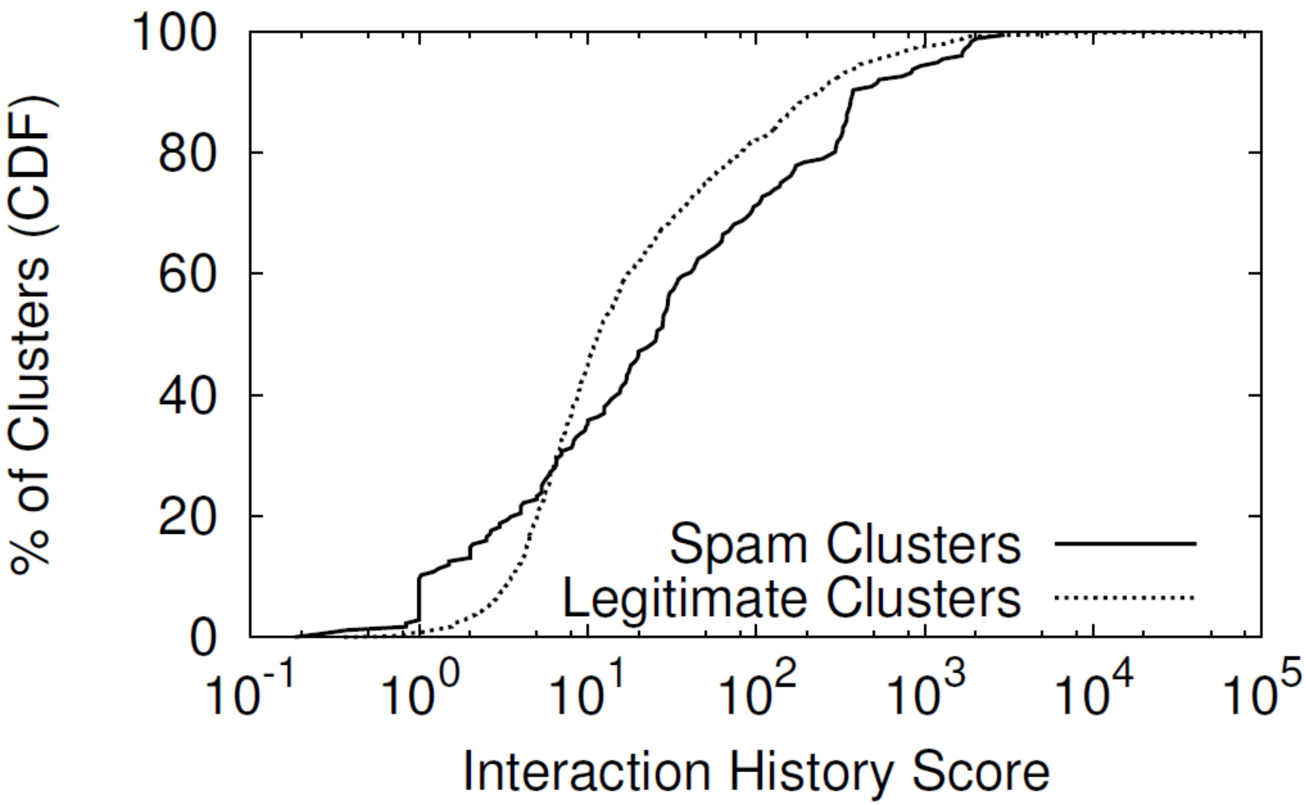


Interaction History

- Legitimate accounts:
  - Normally only interact with a small subset of its friends.
- Spamming accounts:
  - Desire to push spam messages to as many recipients as possible.
- Hypothesis:
  - Spam messages are more likely to be interactions between friends that rarely interact with before.

Interaction History

Interaction history score of spam and legitimate clusters, respectively.



Other Thoughts

- Scalability
  - 300M tweets/day
  - Map-reduce style and cloud computing?