WarningBird: Detecting Suspicious URLs in Twitter Stream

NDSS 2012

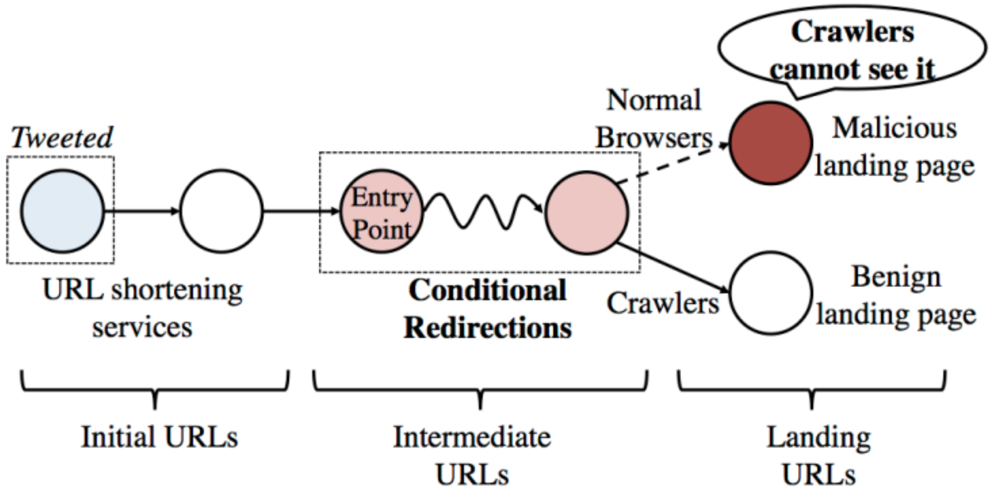
Sangho Lee and Jong Kim
POSTECH, Korea

February 8, 2012

Suspicious URLs in Twitter

- Twitter suffers from malicious tweets.
  - Containing URLs for spam, phishing, …
- Many detection schemes rely on
  - Features of Twitter accounts and msgs.
  - Features of URL and content
- Many evading techniques also exist.
  - Feature fabrication
  - **Conditional redirection**

Conditional Redirection



- Attackers distribute initial URLs of conditional redirect chains via tweets.
- Conditional redirection servers will lead
  - Normal browsers to malicious landing pages
  - **Crawlers to benign landing pages**
    - User agent, IP addresses, repeated visiting, …

Misclassifications can occur

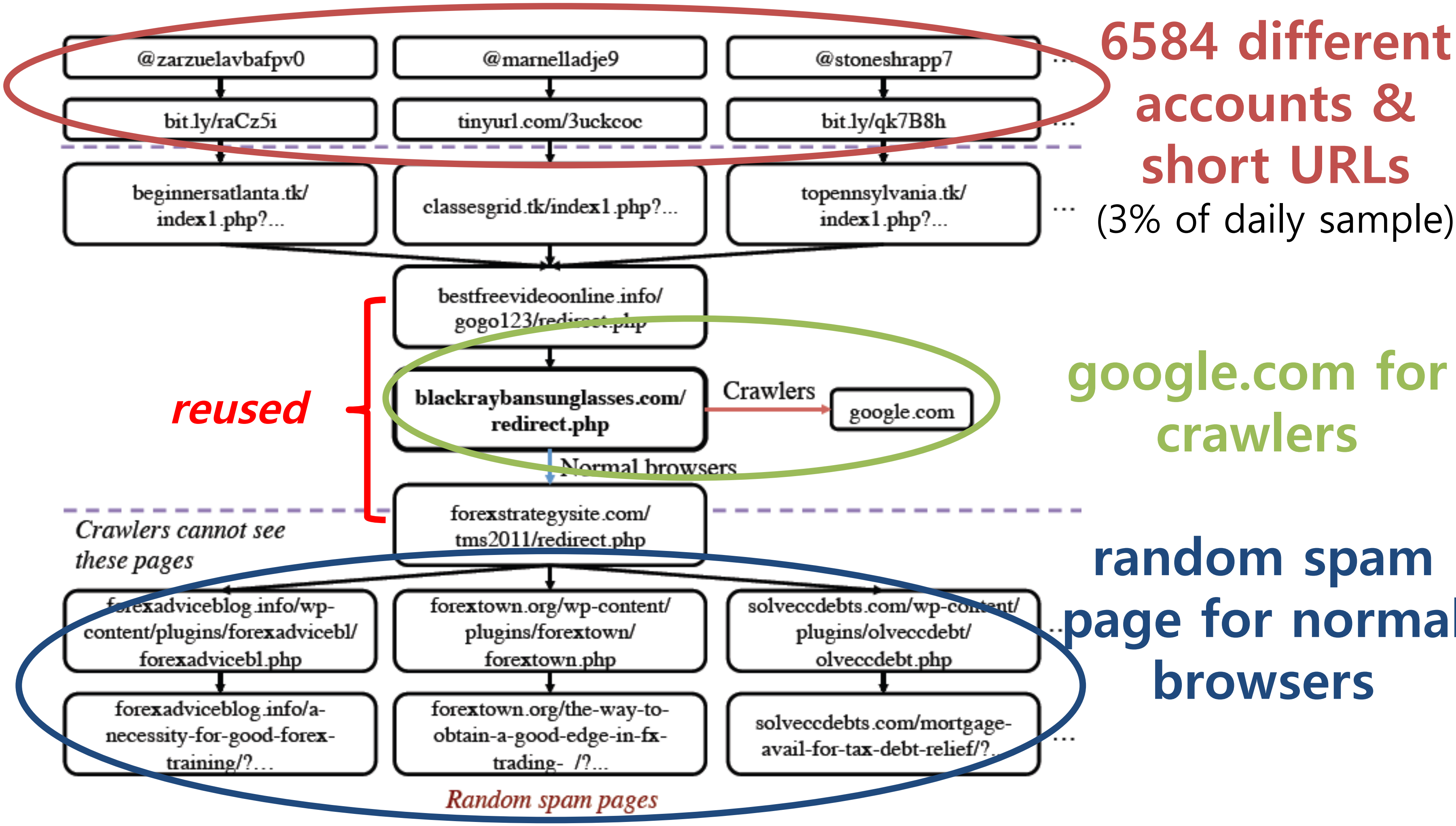
Motivation and Goal

- Attackers can evade previous detection schemes.
  - Selectively provide malicious content to normal browsers not to investigators
- We propose a novel suspicious URL detection system for Twitter.
  - Be robust against evasion techniques
  - Detects suspicious URLs in real time

Outline

- Introduction
- **Case Study**
- Proposed Scheme
- Evaluation
- Discussion and Conclusion

Case Study
blackraybansunglasses.com

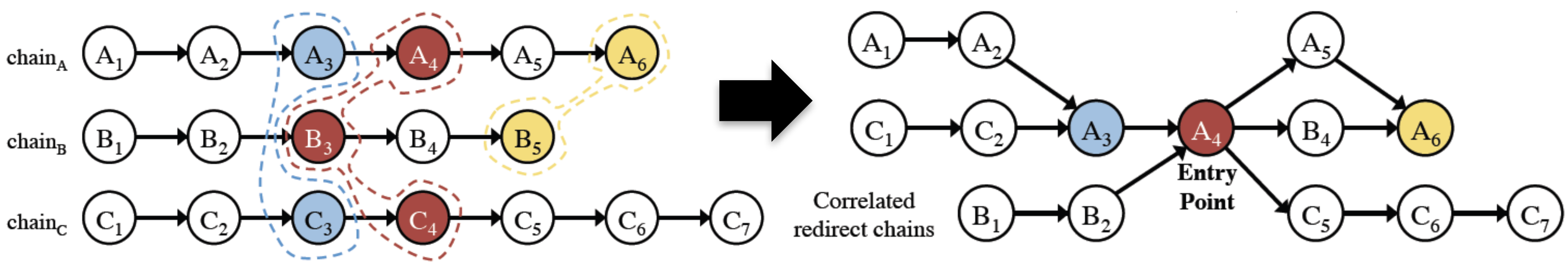


Outline

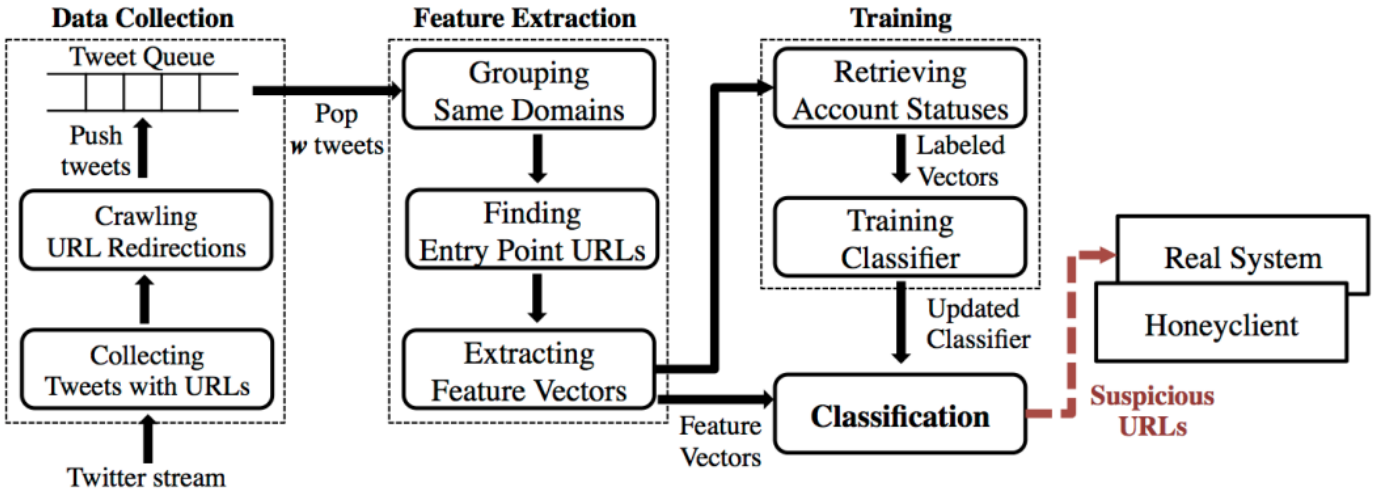
- Introduction
- Case Study
- **Proposed Scheme**
  - **Basic Idea**
  - **System Overview**
  - **Derived Features**
- Evaluation
- Discussion and Conclusion

Basic Idea

- Attackers need to reuse redirection servers.
  - No infinite redirection servers
- We analyze a group of correlated URL chains.
  - To detect redirection servers reused
  - To derive features of the correlated URL chains

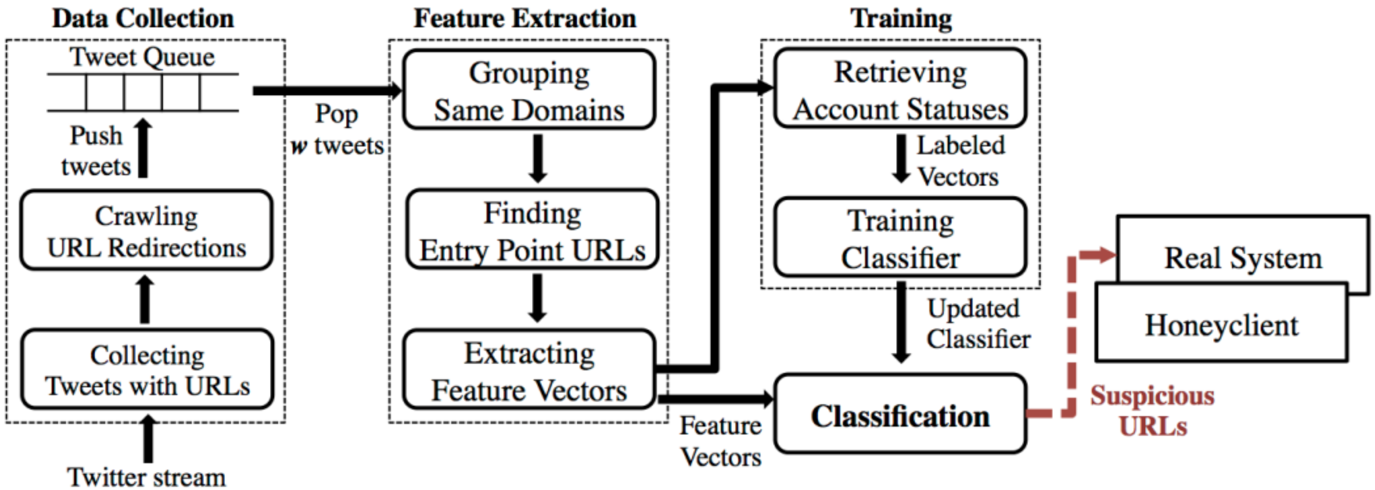


System Overview



- Data collection
  - Collect tweets with URLs from public timeline
  - Visit each URL to obtain URL chains and IP addresses
- Feature extraction
  - Group domains with the same IP addresses
  - Find entry point URLs
  - Generate feature vectors for each entry point

System Overview (continued)



- Training
  - Label feature vectors using account status info.
    - suspended ⇒ malicious, active ⇒ benign
  - Build classification models
- Classification
  - Classify suspicious URLs

Features

- Correlated URL chains
  - Length of URL redirect chain ⬆
  - Frequency of entry point URL ⬆
  - # of different initial and landing URLs ⬆
- Tweet context information
  - # of different Twitter sources ⬇
  - Standard deviation of account creation dates ⬇
  - Standard deviation of friends-followers ratios ⬇

Outline

- Introduction
- Case Study
- Proposed Scheme
- **Evaluation**
  - **System Setup and Data Collection**
  - **Training Classifiers**
  - **Data Analysis**
  - **Detection Efficiency**
  - **Running Time**
- Discussion and Conclusion

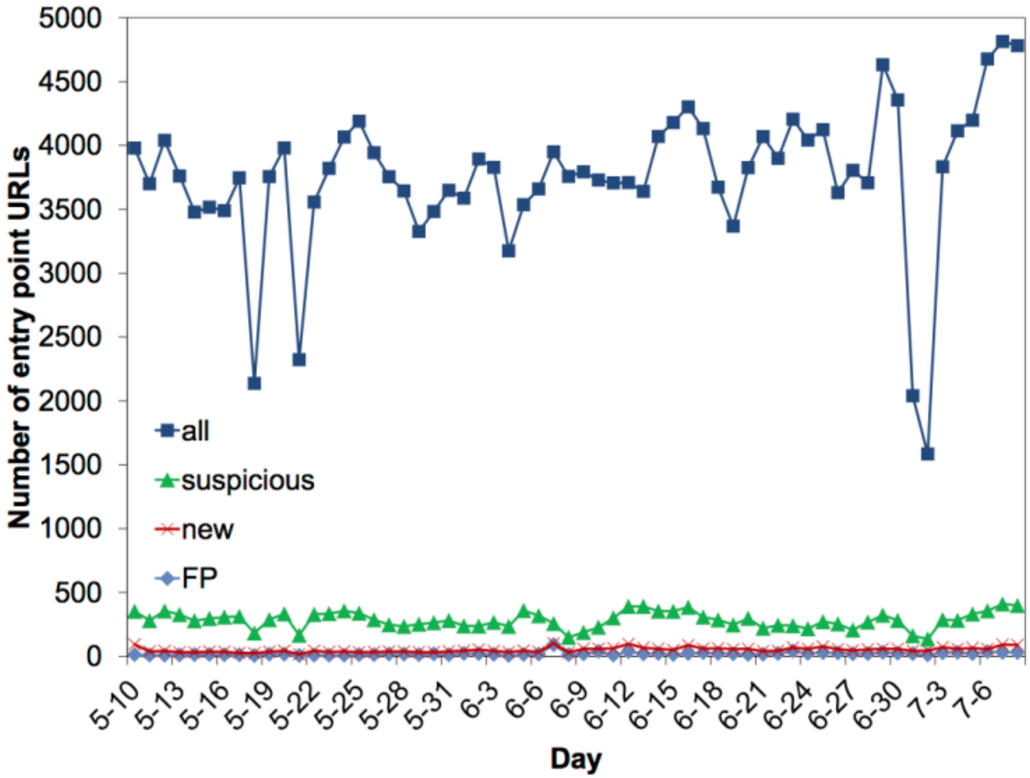
System Setup and Data Collection

- System specification
  - Two Intel Quad Core Xeon 2.4 GHz CPUs
  - 24 GiB main memory
- Data collection
  - Twitter Streaming API
  - One percent samples from Twitter public timeline (Spritzer role)
  - 27,895,714 tweets with URLs between April 8 and August 8, 2011 (122 days)

Training Classifiers

- Training dataset
  - Tweets between May 10 and July 8
  - 183,113 benign and 41,721 malicious entry point URLs
- Classification algorithm
  - L2-regularized logistic regression
- 10-fold cross validation
  - FP: 1.64%, FN: 10.69%

Data Analysis

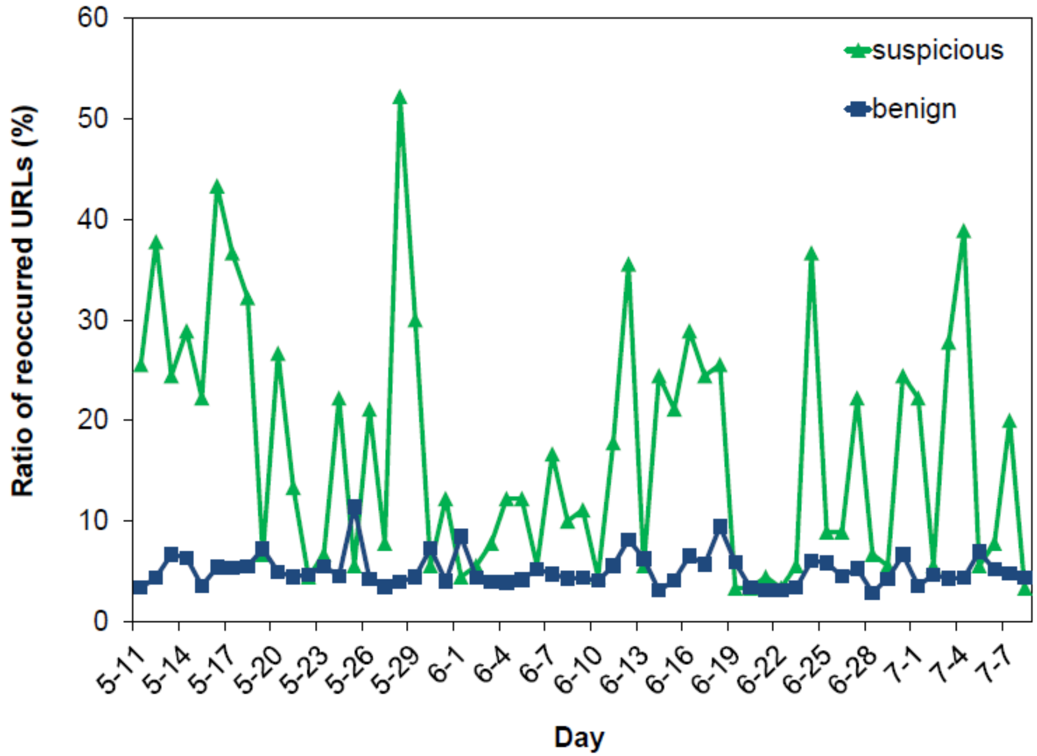


3758 entry point URLs
(on average, daily)

283 suspicious URLs
20 false positive URLs
30 new suspicious URLs

- Relatively small number of new suspicious URLs
  - We detect suspicious URLs that are not detected or blocked by Twitter.

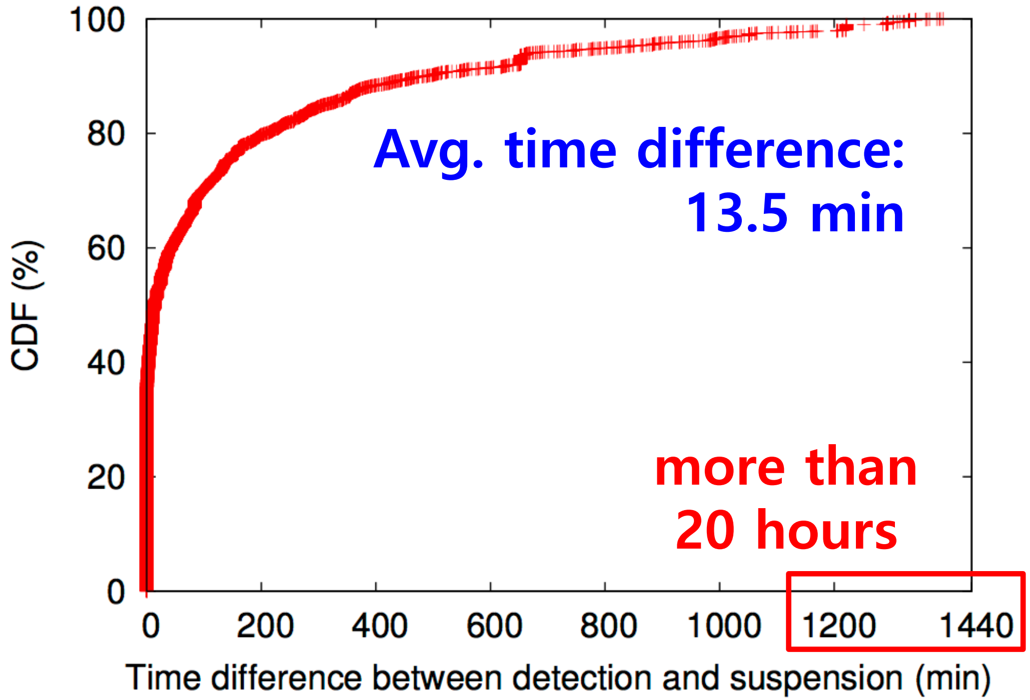
Data Analysis (continued)



- Reoccurrences of May 10's URLs
  - Up to 12% benign & 52% suspicious URLs

Detection Efficiency

- We measure the time difference between
  - When WarningBird detects suspicious accounts
  - When Twitter suspends the accounts



Running Time

- Processing time for each URL: 28.31 ms
  - Redirect chain crawling: 24.20 ms
    - Hundred crawling threads
  - Domain grouping: 2.00 ms
  - Feature extraction: 1.62 ms
  - Classification: 0.48 ms
- Our system can classify about 127,000 URLs per hour.
  - About 12.7% of all public tweets with URLs per hour

Outline

- Introduction
- Case Study
- Proposed Scheme
- Evaluation
- **Discussion and Conclusion**

Discussion

- Evasion is possible but restricted.
  - Do not reuse redirection servers
    - Need extra $ (to buy compromised hosts)
    - Need more effort to take down hosts
  - Reduce the rate of malicious tweets
    - Less effective

Conclusion

- We proposed a new suspicious URL detection system for Twitter.
- Our system is robust against feature fabrication and conditional redirection.
- Evaluation results show accuracy and efficiency.