Spam Research

Yue Duan
Illinois Institute of Technology

@spam: The Underground on 140 characters or less

ChrisGrier, KurtThomas, VernPaxson and MichaelZhang

University of California, Berkeley

CCS 2010

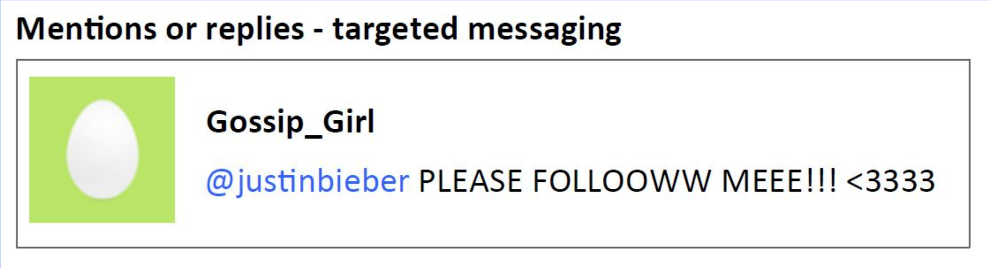
Background

- Spam on Twitter
  - goal: in-depth understanding of spam on Twitter
- Twitter is social network and messaging app
  - Over 190 million visitors per month
  - Over 2 billion messages per month
- Social networks a major target for spammers
  - 10% of URLs posted on Facebook lead to spam

Background

- Characterization of spam on Twitter
    - Use of social features
    - Specific campaigns
- We found 3 million tweets containing spam URLs
    - 8% of URLs posted lead to spam content
    - Collection lasted one month, in January 2010
- Directly measure click--through, determine success

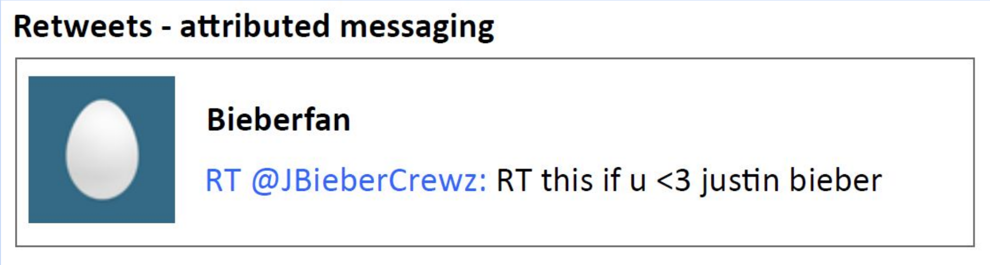
Background



Mentions or replies - targeted messaging

Gossip_Girl

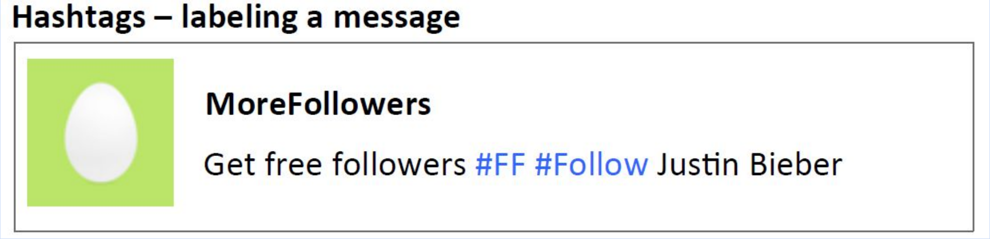
@justinbieber PLEASE FOLLOOWW MEEE!!! <3333



Retweets - attributed messaging

Bieberfan

RT @JBieberCrewz: RT this if u <3 justin bieber



Hashtags – labeling a message

MoreFollowers

Get free followers #FF #Follow Justin Bieber

Background

- RT@scammer:check out the ipads there having a give-away http://spam.com
- Buy more followers! http://spam.com#fwlr
- http://spam.com RT@barackobama A great battle is ahead of us
- Help donate to #haiti relief:http://spam.com

Collecting Tweets

- Use publicly available Twitter APIs
  - Streaming and REST APIs
- 200+ million Tweets with URLs from stream
  - Jan--Feb 2010, one month of collection
- 150k users their complete history
  - Randomly sampled users from stream
  - 200+ million Tweet

Classifying Tweets

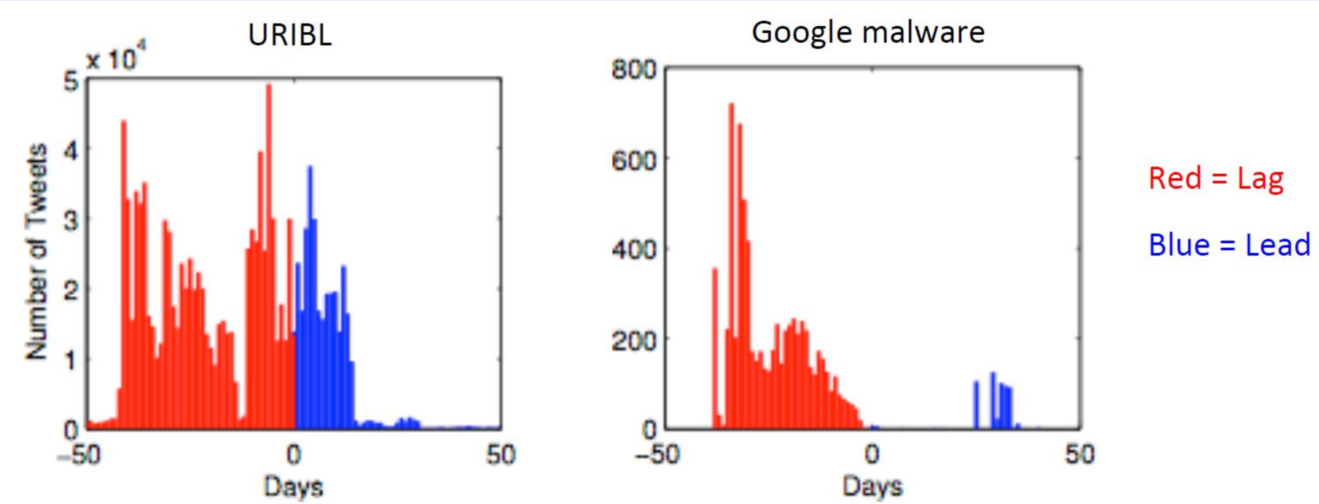
- Only concerned with Tweets containing URLs
- Classifying Tweets
  - Manual classification - 26% of URLs lead to spam
    - Use a browser, click on the URL, classify as spam or ham
    - 5% error at 95% confidence
  - Automatic – 8% of URLs lead to spam
    - Use existing domain and URL blacklists
    - Google Safebrowsing : malware + phishing
    - URIBL : email spam
    - Joewein : email spam

Blacklisting URLs

- Over 80% of spam URLs were shortened
    - Need the final URL or landing site to blacklist
    - Mask landing site
        - http://bit.ly/aLEmck --> htpp://i--drugspedia.com/pill/Viagra...
    - Defeat blacklist filtering
        - bit.ly --> short.to --> malware landing page
- Crawl URLs to find landing site
    - 25 million URLs crawled

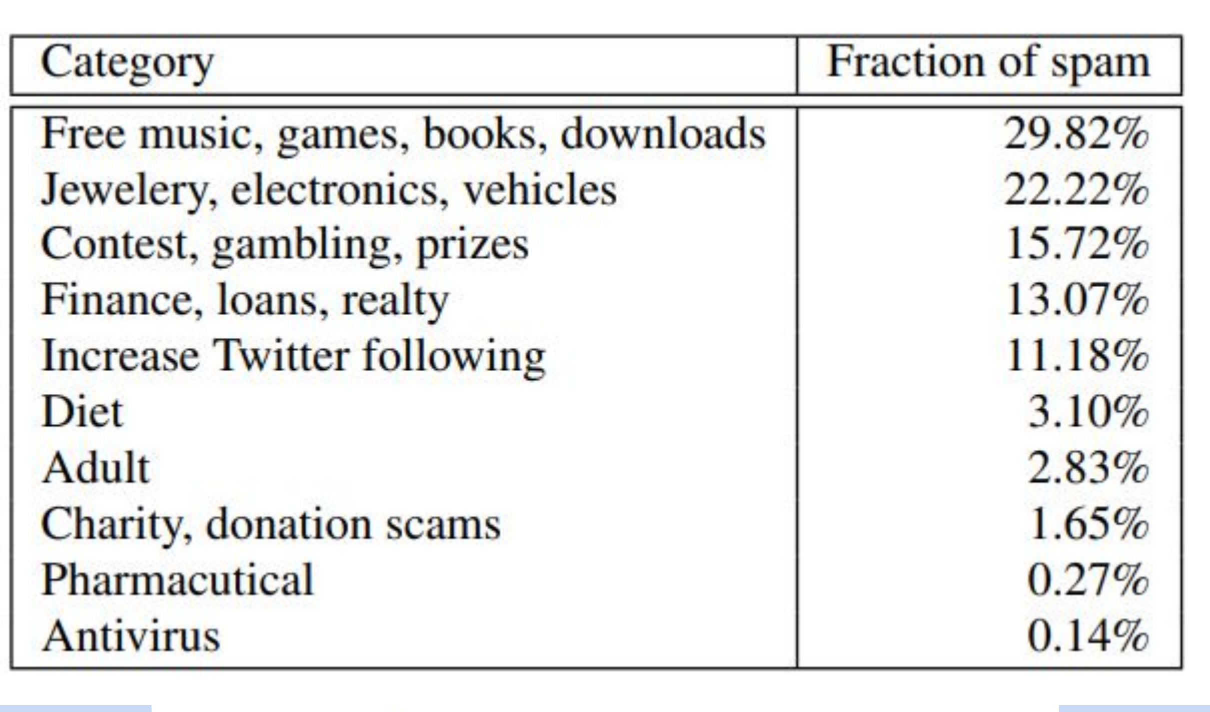
Blacklist Performance

- Blacklists are slow to list spam domains
  - 80% of clicks are seen in first day
- Retroactively blacklist



Spam Statistics

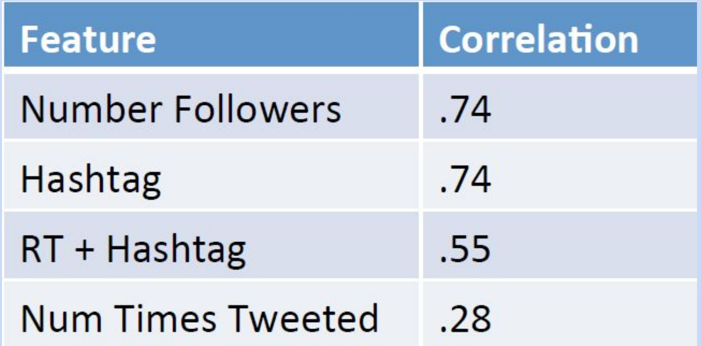
- Crawled
  - 2 mi
  - 3 mi



Category	Fraction of spam
Free music, games, books, downloads	29.82%
Jewelery, electronics, vehicles	22.22%
Contest, gambling, prizes	15.72%
Finance, loans, realty	13.07%
Increase Twitter following	11.18%
Diet	3.10%
Adult	2.83%
Charity, donation scams	1.65%
Pharmacutical	0.27%
Antivirus	0.14%

Spam Statistics

- Spam Clickthrough
  - 245,000 spam URLs with clickthrough stats
    - 97.7% receive 0 clicks
    - 2.3% receive over 1.6 million clicks
  - Successful spam Tweets
    - Linear correlation between clicks and features



Feature	Correlation
Number Followers	.74
Hashtag	.74
RT + Hashtag	.55
Num Times Tweeted	.28

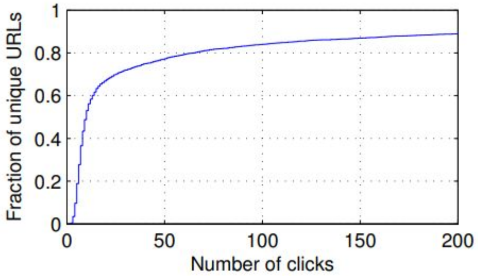


Figure 1: Clickthrough for spam URLs posted to Twitter. Only the 2.3% of URLs that generated any traffic are shown.

Comparison to Email Clickthrough

- Spam Email clickthrough: .003--.006%
  - From SpamalyUcs, Kanich et al. CCS 2008
- Twitter clickthrough: .13%
  - Define clickthrough as clicks / reach
  - Reach defined as *tweets* * *followers*

Spamming Accounts

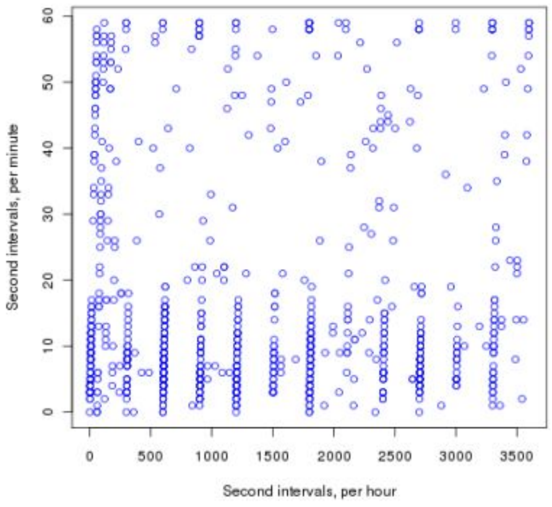
- Are accounts being created to spam?
  - “career” spammers
- Accounts being compromised for spam?
- Two tests to determine account state
  - χ 2 test on tweet timestamps
    - Seconds of the minute
    - Seconds of the hour
  - Text entropy
    - Same text
    - Same URL

Spamming Accounts

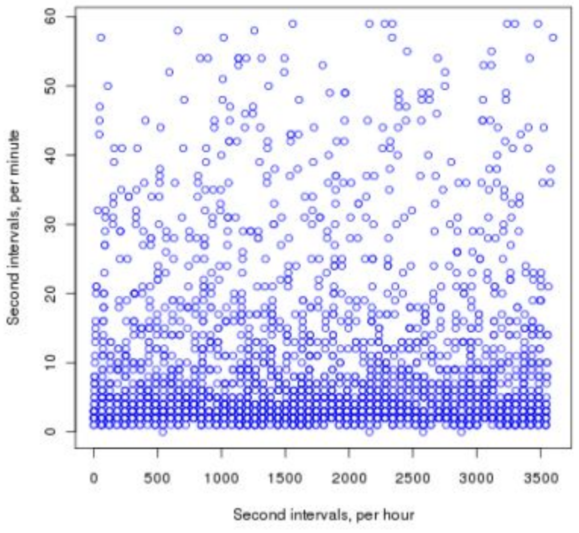
- χ 2 test on tweet timestamps
  - assumption: legitimate account tweets overall reflect a **uniform** process
  - examines tweet timestamps to identify patterns in the minutes and seconds for when a tweet was posted
  - represent timestamps for an individual account using vectors corresponding to the **seconds value of each hour** and **seconds value of each minute**
  - compute the **p-value** for these vectors for their consistency with an underlying uniform distribution
  - e.g., p-value < 0.001 indicates less than 0.1% chance that a user posting as a Poisson process generated the sequence.

Spamming Accounts

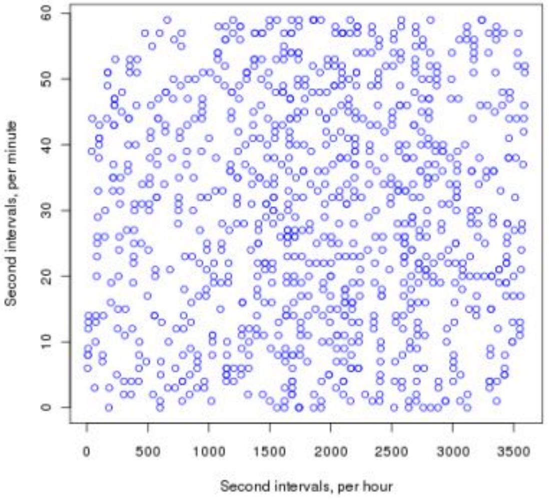
- χ 2 test on tweet timestamps



(a)



(b)

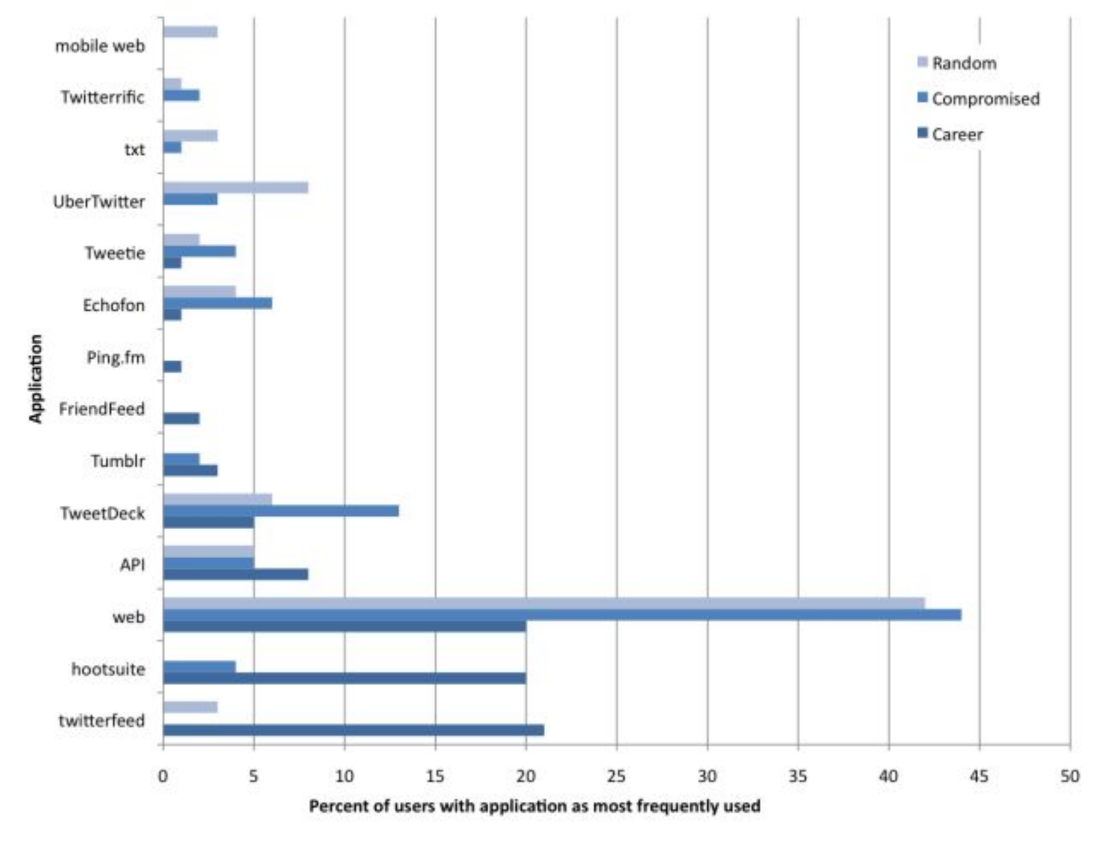


(c)

Compromised Accounts

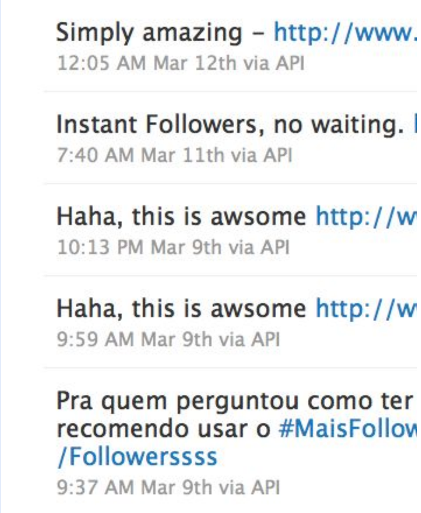
- The majority of accounts pass both tests
    - Accounts are being stolen for spam use
        - Phishing, password guessing
        - Malware using Twitter accounts
- Compromised account evidence
    - Application Use
        - 22% of accounts contain spam tweets from applications never used for non-spam tweets.
    - Setup a fake account as a spam trap
        - Provided credentials to a frequently tweeted phishing site
        - Account then used to advertise phishing and other scams
        - Over 20,000 other users had tweeted same links

Compromised Accounts



Compromised Accounts

- Compromised account evidence (cont.)
  - Infiltrated Koobface and identified Koobface tweet templates
    - Koobface is a botnet th
    - Stolen accounts tweet



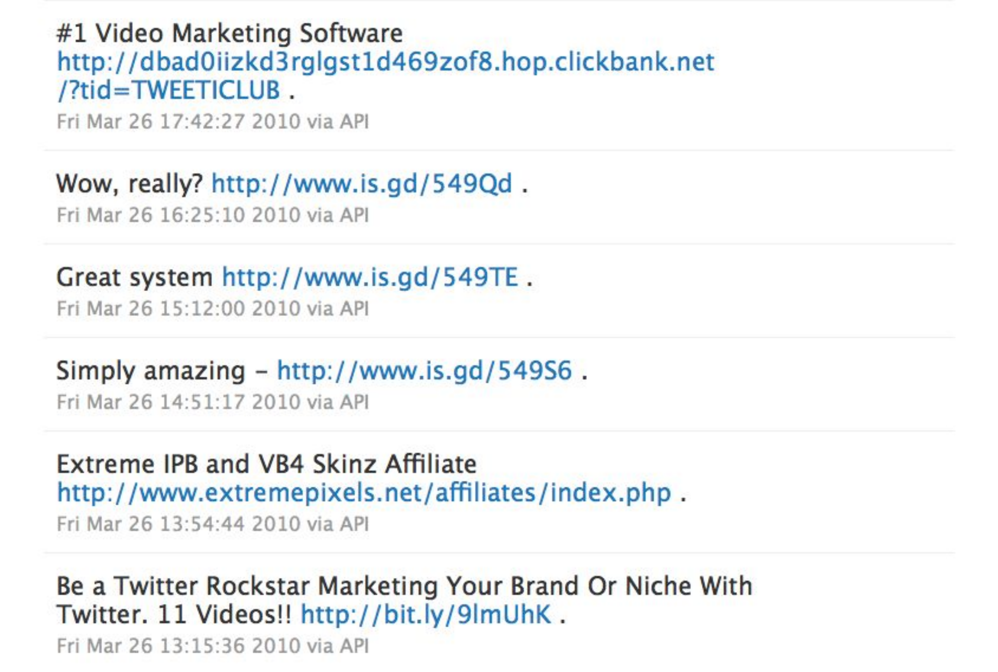
Simply amazing – http://www.
12:05 AM Mar 12th via API

Instant Followers, no waiting. h
7:40 AM Mar 11th via API

Haha, this is awsome http://w
10:13 PM Mar 9th via API

Haha, this is awsome http://w
9:59 AM Mar 9th via API

Pra quem perguntou como ter
recomendo usar o #MaisFollow
/Followerssss
9:37 AM Mar 9th via API



#1 Video Marketing Software
http://dbad0iizkd3rglgst1d469zof8.hop.clickbank.net/?tid=TWEETICLUB .
Fri Mar 26 17:42:27 2010 via API

Wow, really? http://www.is.gd/549Qd .
Fri Mar 26 16:25:10 2010 via API

Great system http://www.is.gd/549TE .
Fri Mar 26 15:12:00 2010 via API

Simply amazing – http://www.is.gd/549S6 .
Fri Mar 26 14:51:17 2010 via API

Extreme IPB and VB4 Skinz Affiliate
http://www.extremepixels.net/affiliates/index.php .
Fri Mar 26 13:54:44 2010 via API

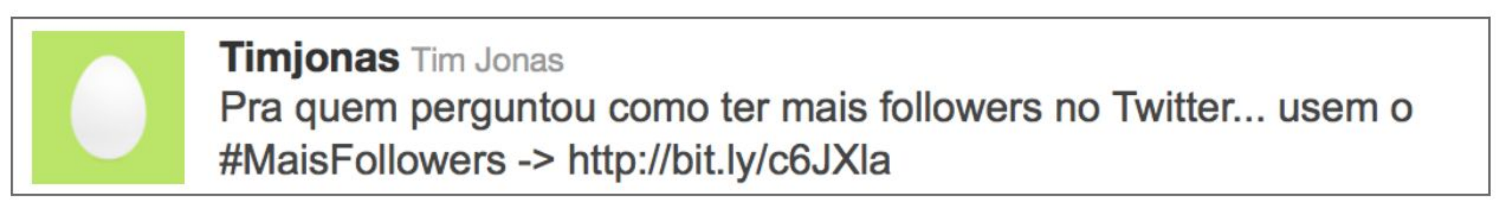
Be a Twitter Rockstar Marketing Your Brand Or Niche With Twitter. 11 Videos!! http://bit.ly/9lmUhK .
Fri Mar 26 13:15:36 2010 via API

Spam Campaigns

- Cluster URLs to find campaigns
  - Cluster defined by a binary feature vector {0,1}n
  - n is the total number of spam URLs
  - Merge clusters with URLs in common
- Limitations
  - Merges campaigns if users participate in multiple campaigns
  - Will not merge if users do not share URLs

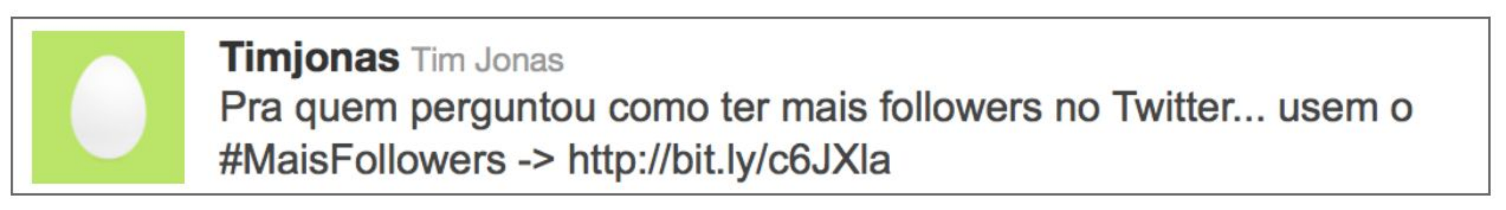
Campaign: Phishing for Followers

- Clustering found 1,120 different URLs
  - Posted by 21,284 users
  - Leading to 12 different domains
  - URLs contained affiliate IDs
- Defining characteristics
  - 88% of users were compromised users
  - Extensive use of similar hashtags
  - Two hop redirect chain: short --> affiliate link --> landing site



Timjonas Tim Jonas
Pra quem perguntou como ter mais followers no Twitter... usem o #MaisFollowers -> http://bit.ly/c6JXla

Campaign: Phishing for Followers



Timjonas Tim Jonas

Pra quem perguntou como ter mais followers no Twitter... usem o #MaisFollowers -> http://bit.ly/c6JXla

Rough translation : “ For those who asked for more followers on Twitter...Use#MaisFollowers”

Conclusion

- Spam on Twitter is abundant and successful
  - 26% of URLs lead to spam
  - Clickthrough over 10x that of email spam
- Spammers are compromising accounts for use
  - Require accounts to send spam
- Adopting social elements for use in spam
  - URL shortening to mask destination, evade blacklists
  - Hashtags, retweets, correlated with successful spam