

@spam: The Underground on 140 characters or less

Chris Grier, Kurt Thomas,
Vern Paxson and Michael Zhang
University of California, Berkeley



Why 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
 - Twitter XSS used to redirect visitors to porn

Twitter Spam

- 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

Outline


- Twitter background
- Collection and classification
- Spam analysis
 - Click-through for spam
 - Finding compromised accounts
 - Spam Campaigns
- Conclusion



- Simple messages called Tweets
 - 140 characters or less
- A user has *followers* and *friends*
 - Relationships are not bi-directional
 - Alice's *followers* see the tweets sent by Alice
 - Alice sees tweets that her *friends* post

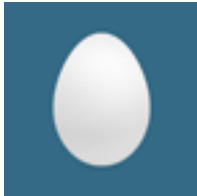
Components of a Tweet

Mentions or replies - targeted messaging



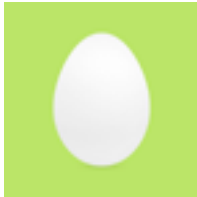
Gossip_Girl
[@justinbieber](#) PLEASE FOLLOOWW MEEE!!! <3333

Retweets - attributed messaging



Bieberfan
RT [@JBieberCrewz](#): RT this if u <3 justin beiber

Hashtags – labeling a message



MoreFollowers
Get free followers [#FF](#) [#Follow](#) Justin Bieber

Spammers Tweeting?

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 Tweets

Classifying Tweets

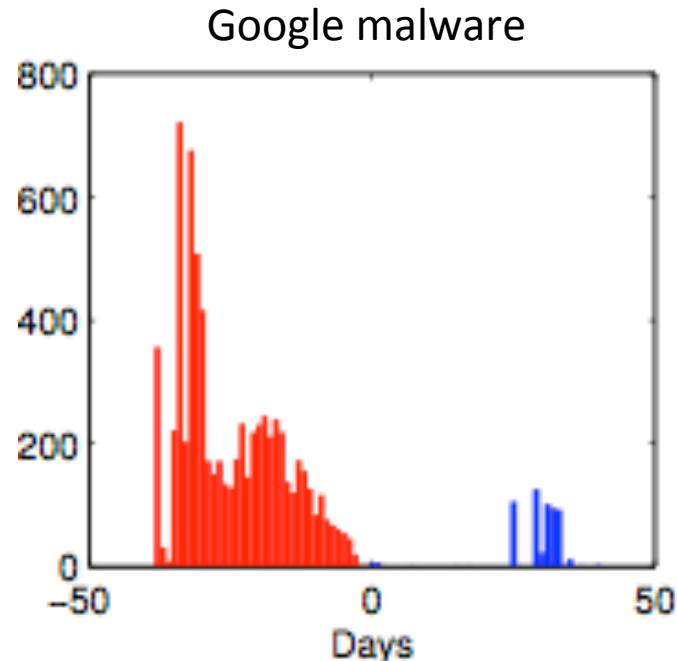
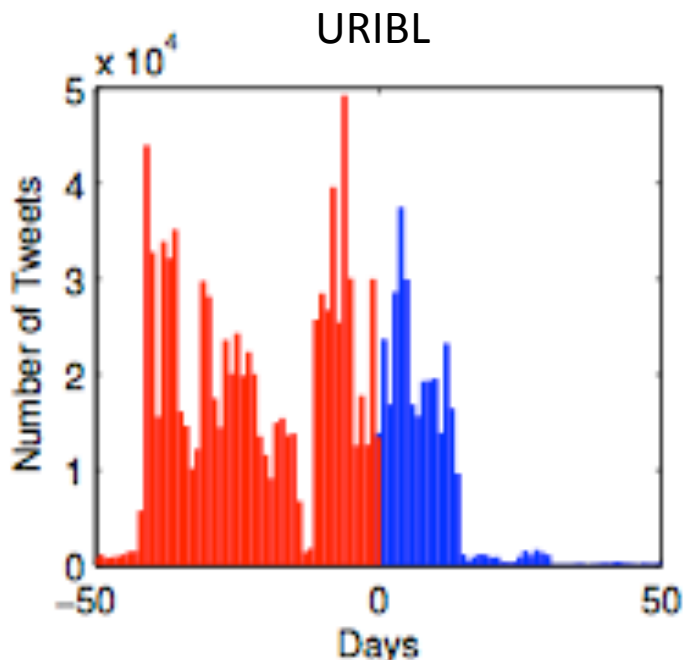
- Only concerned with Tweets containing URLs
- Classifying Tweets
 - Manual classification – 26% of URLs lead to spam
 - Use a browser, click on the URL, classify as spam or ham.
 - 5% error at 95% confidence
 - Automatic – 8% of URLs lead to spam
 - Use existing domain and URL blacklists
 - Google Safebrowsing : 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> -> <http://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



Red = Lag

Blue = Lead

Spam Analysis

General Spam Statistics

- Crawled 25 million URLs and checked blacklists
 - 2 million lead to *known* spam, phishing and malware
 - 3 million tweets contained a spam URL

	#	@	RT @
Tweets	13.3%	41.1%	13.6%
Tweets w/URL	22.4%	14.1%	16.9%

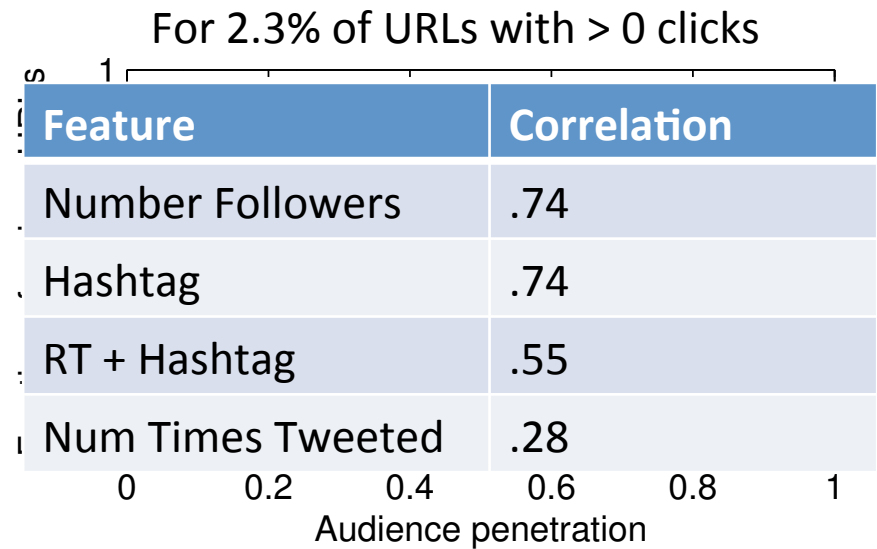
Global use of Twitter features

Google Safebrowsing	70.1%	3.5%	1.8%
Joewein	5.5%	3.7%	6.5%
URIBL	18.2%	10.6%	11.4%

Spam use of Twitter Features

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



Comparison to Email Clickthrough

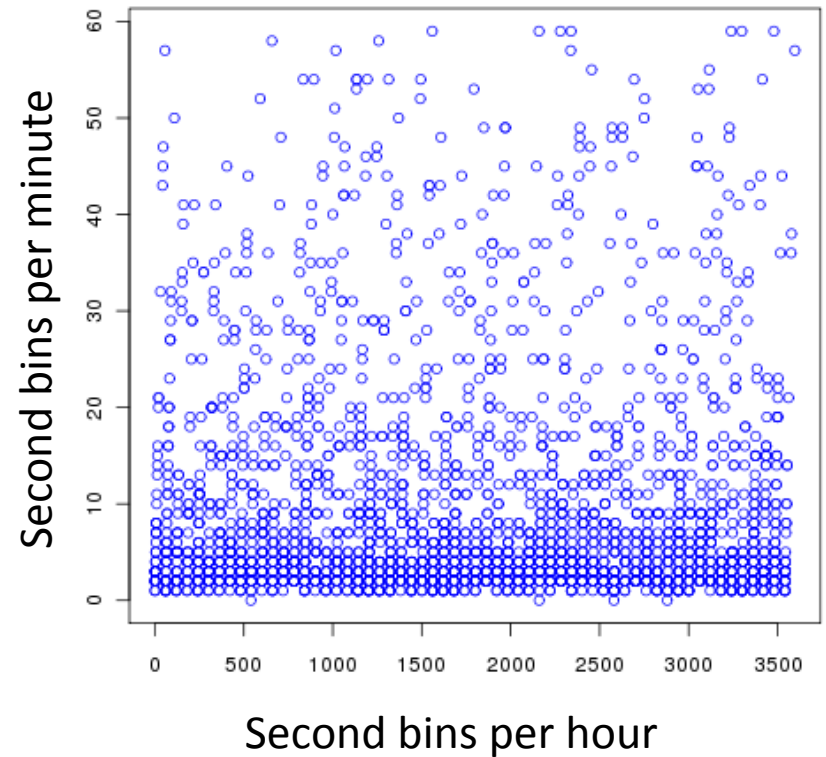
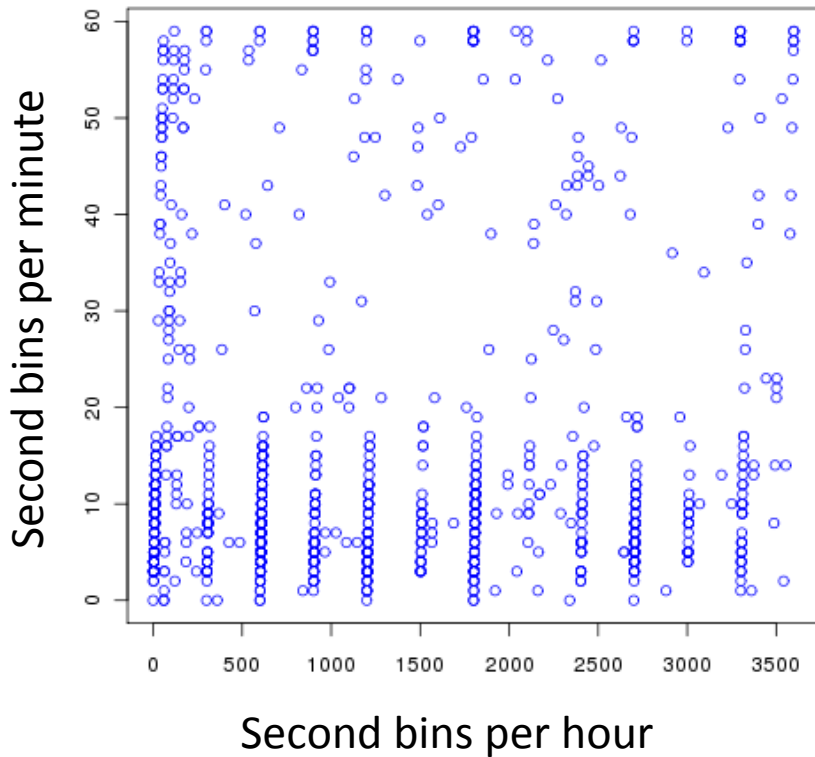
- Spam Email clickthrough: .003-.006%
 - From Spamalytics, Kanich et al. CCS 2008
- Twitter clickthrough: .13%
 - Define clickthrough as clicks / reach
 - Reach defined as *tweets * followers*

Spamming Accounts

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

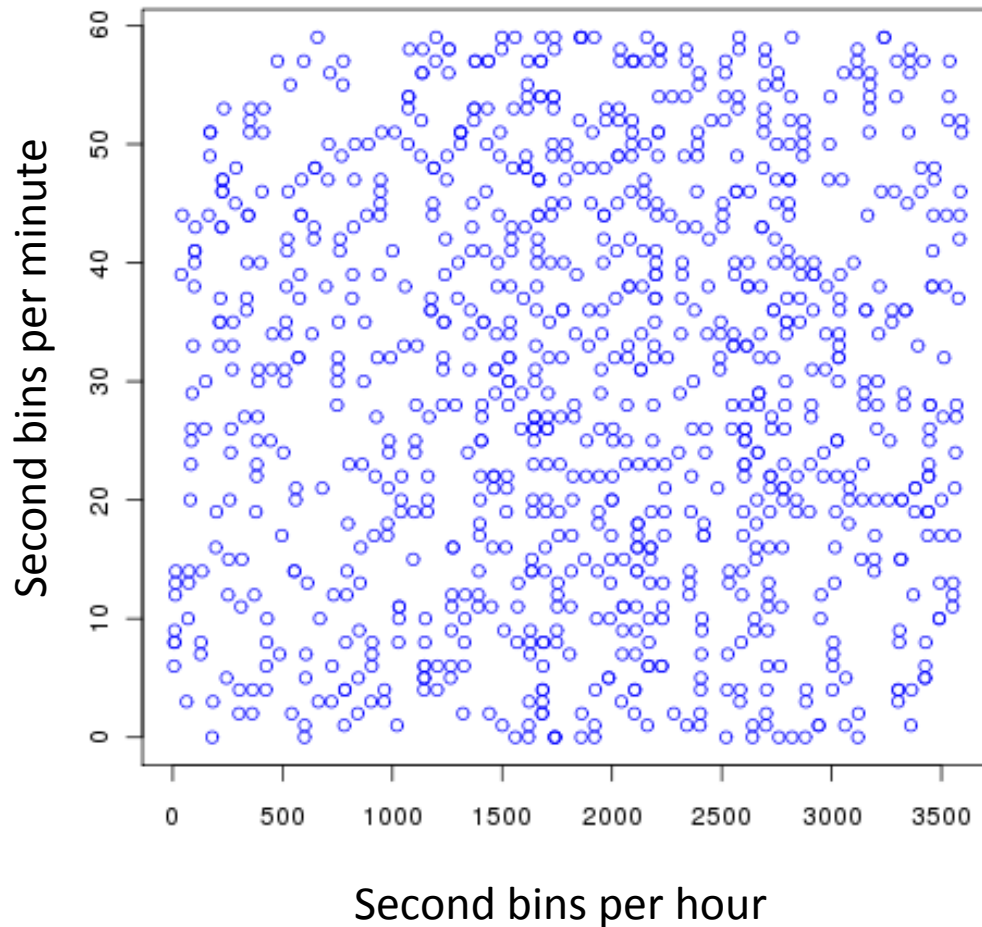
χ^2 Test on Timestamps

Tweets binned by time for a two users
Patterns indicate regular posting intervals



χ^2 Test on Timestamps

Tweets binned by time for a third user



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
 - Infiltrated Koobface and identified Koobface tweet templates
 - Koobface is a botnet that spreads via social networks, steals credentials
 - Stolen accounts tweet links to Koobface installer

#1 Video Marketing Software

Simply amazing – <http://www.is.gd/549S6> .

12:05 AM Mar 12th via API

Instant Followers, no waiting. <http://www.is.gd/549Rv> .

7:40 AM Mar 11th via API

Haha, this is awesome <http://www.is.gd/549TE> .

10:13 PM Mar 9th via API

Haha, this is awesome <http://www.is.gd/549TE> .

9:59 AM Mar 9th via API

Pra quem perguntou como ter mais followers no Twitter, recomendo usar o #MaisFollowers -> <http://tinyurl.com/Followerssss>

9:37 AM Mar 9th via API

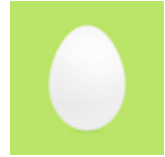
Be a Twitter Rockstar Marketing Your Brand Or Niche With Twitter. 11 Videos!! <http://bit.ly/9ImUhK> .

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



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”
- 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

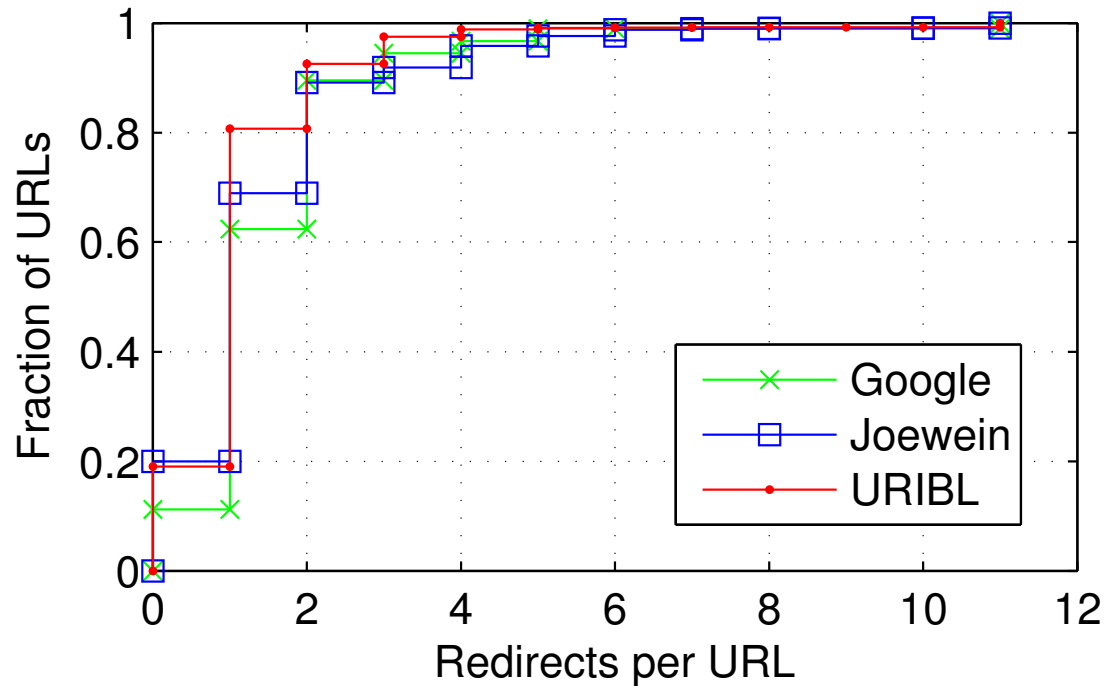
Conclusion

- 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

Questions?

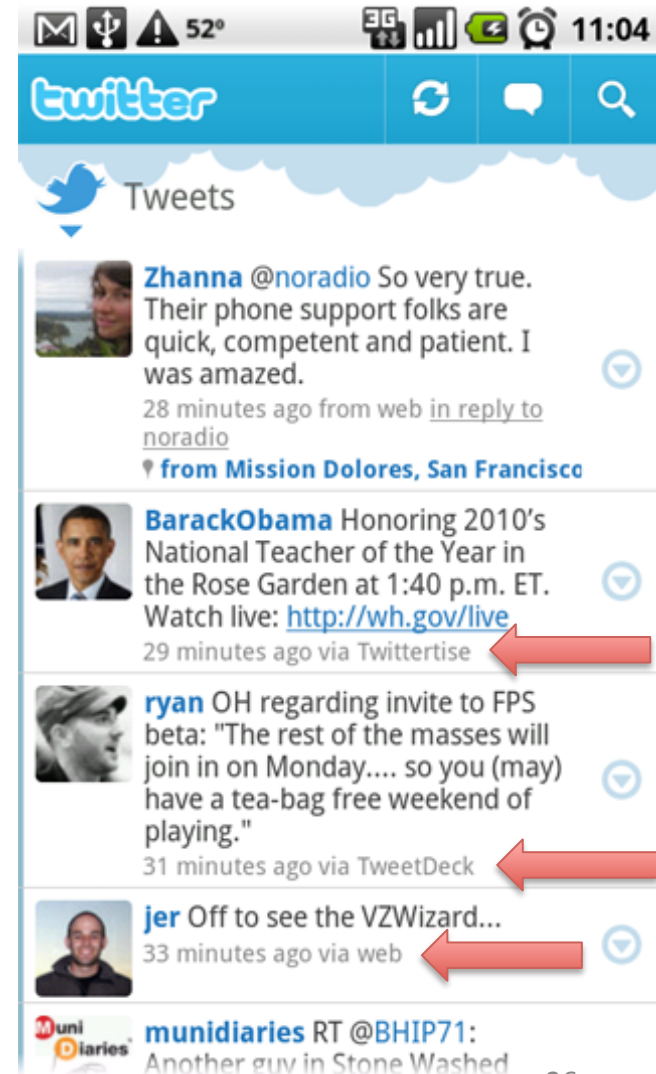
Redirecting to spam

- Follow redirects
 - HTTP and META tag
- 25 million URLs
 - 33 million redirects
- 8 million spam URLs

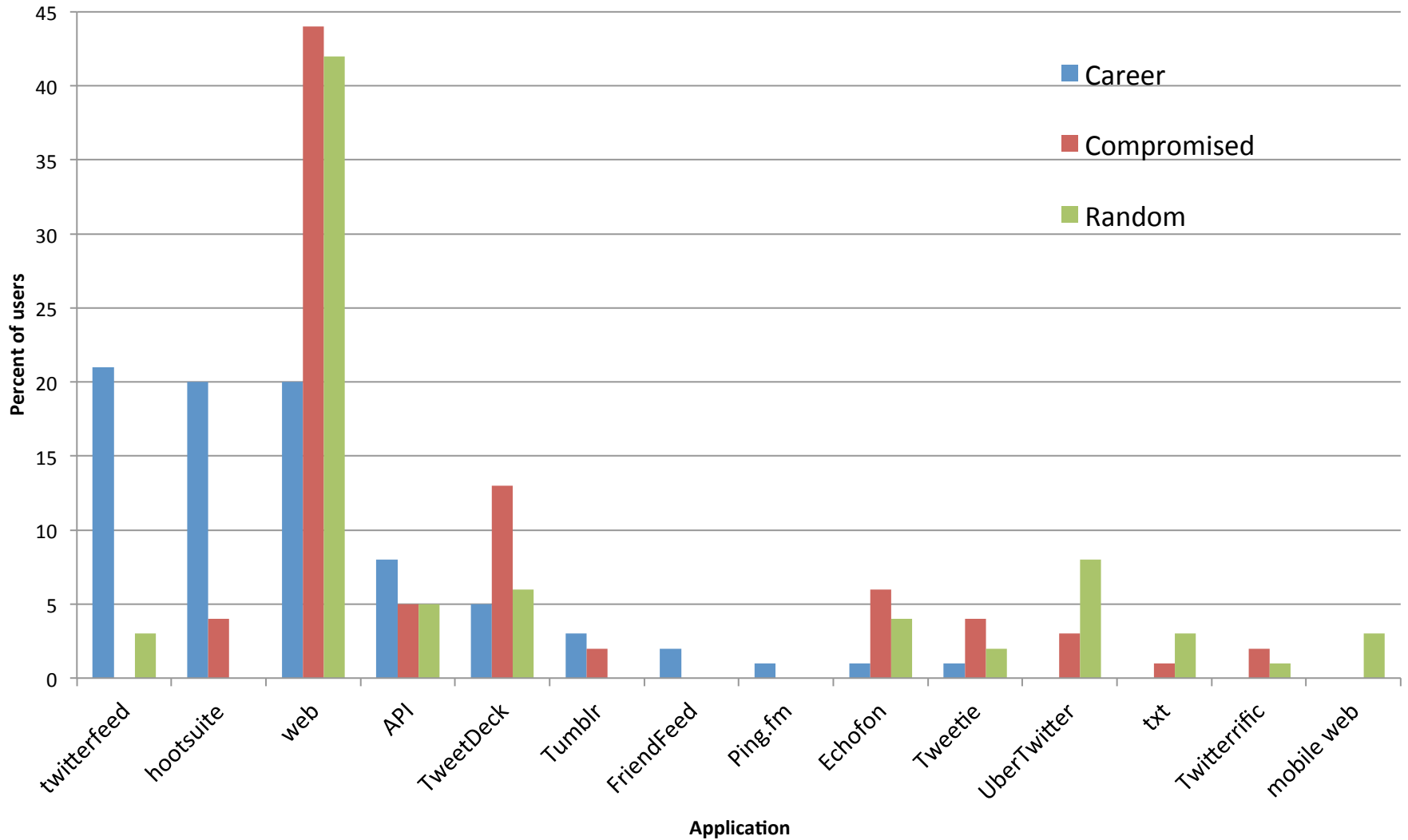


Application use

- Twitter enables third party clients using a REST API
- Platform specific
 - TweetDeck
 - Iphone, Andriod app
- Web based apps
 - Hootsuite, Twitterfeed
- Automatic syndication
- Do spammers use apps?



Applications used by spammers



High level results

- Of 25 million URLs, 2 million lead to *known* malware, phishing and scams
- Analyzed over 500k URLs with click-through data
 - Over 1.6 million clicks on spam URLs
- Up to 80% of the users that sent spam had their account *compromised*
- Blacklists are too slow
 - Malware URLs added 25 days after link posted on Twitter
 - URLs on twitter receive 80% clicks within first day

Other results...

- Blacklists too slow to be effective at filtering
 - Google malware: 25 days
 - Google phishing: 9 days
 - URIBL, Joewein: 22 days, 4 days
- Twitter is a successful platform for spamming users
 - Clickthrough highly correlated with followers
 - 10-100x more successful at getting clicks than email spam

Related Work