

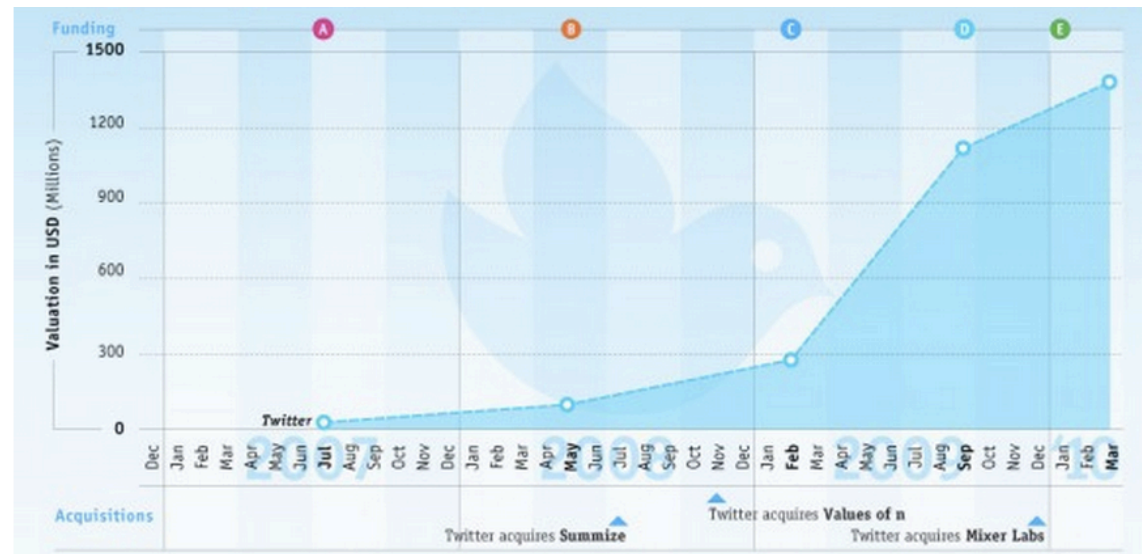
Identifying Trending #SPAM in Twitter

Despoina Antonakaki
despoina@ics.forth.gr

Twitter in numbers

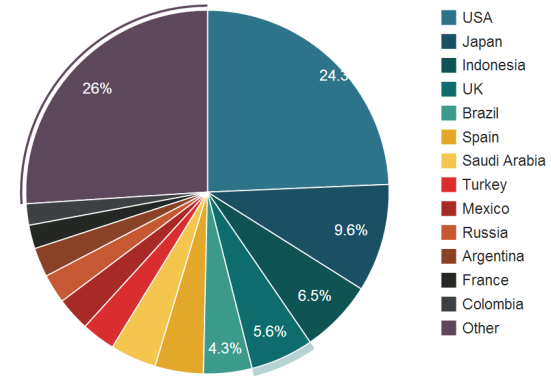
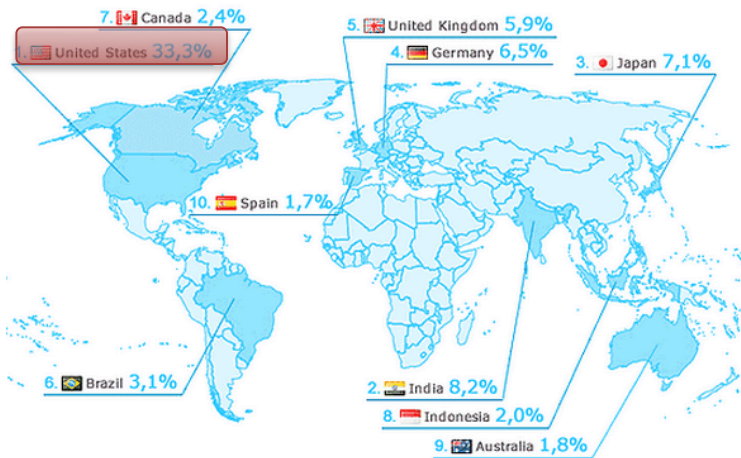
Fast growing OSN:

- Registered users: 650 million,
- 232 million monthly active users,
- Average number of daily tweets: 58 million
- Twitter queries per day: 2.1 billion

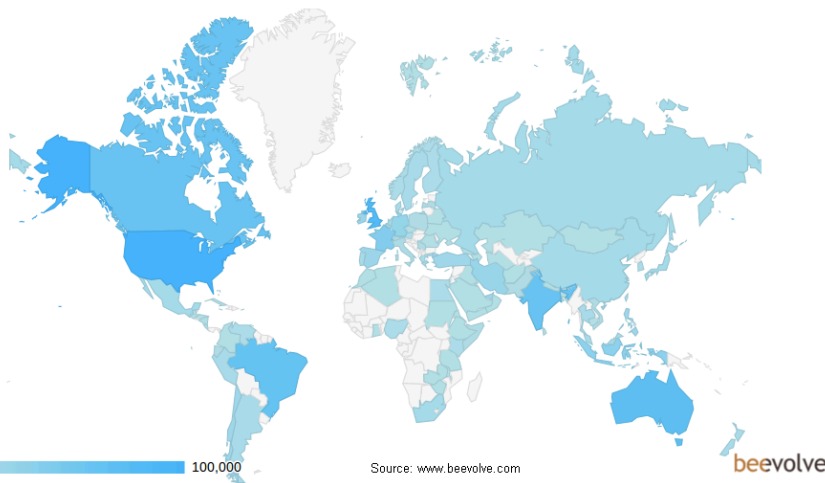


In 2010 - 2012 ...

Top 10 countries (percent of site traffic)



source: peerreach.com



COUNTRY	% OF USERS
United States	50.99
United Kingdom	17.09
Australia	4.09
Brazil	3.44
Canada	2.92
India	2.87
France	1.76
Indonesia	1.43
Iran	0.88
Ireland	0.85

5 out of 10 twitter users are from USA Tweet This →

Popular Trends (PTs)

PTs:

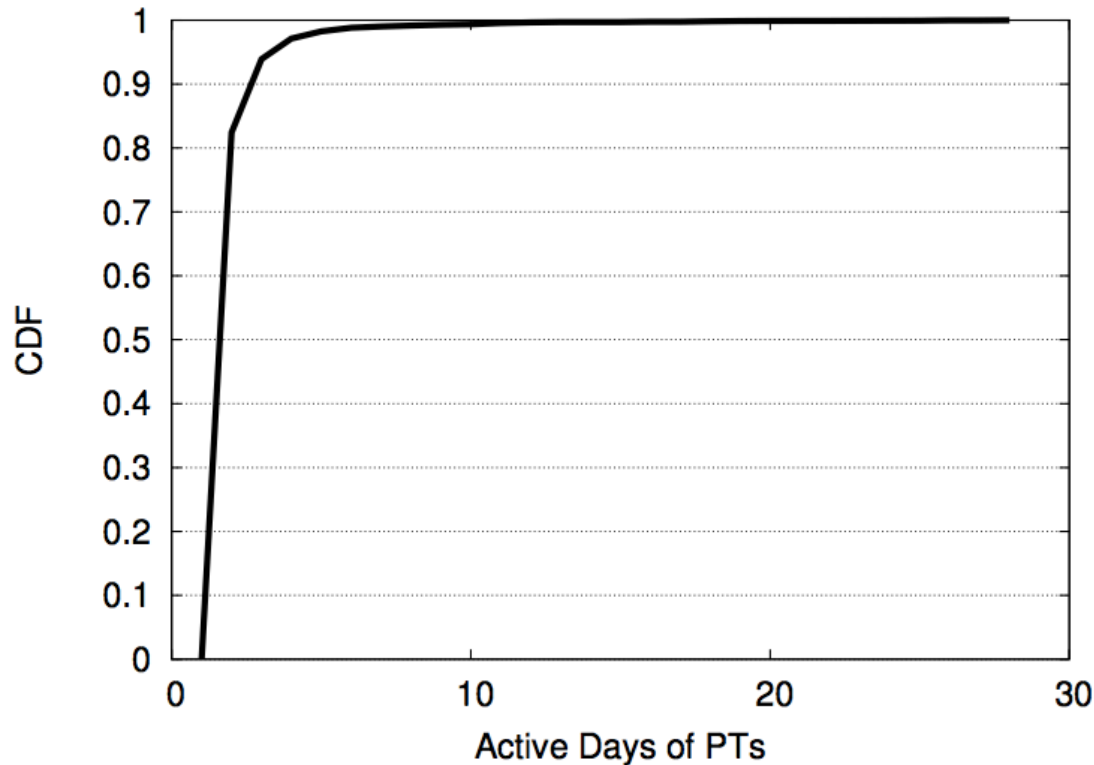
- Introduced in 2010: Hashtags rapidly becoming popular
- Popular #Hashtags
- Common Search Queries
- PTs belong to geographic regions
- Accesses from Twitter's API
- Exploited by Spammers:



The Dataset:

- **January – February – March 2014**
- Request 10 Popular Trends (PTs) every 10 minutes.
 - In average 240 PTs per day
- **10 Twitter accounts**
 - Each account grabs a trend and downloads 1000 tweets
 - Then proceed to next trend
- In average: 1.5M-2.0M tweets / day
- **Data from US**
- In average: 400.000 different users / day
- Total 150M tweets with PTs

How long does a PT “stays alive” ?



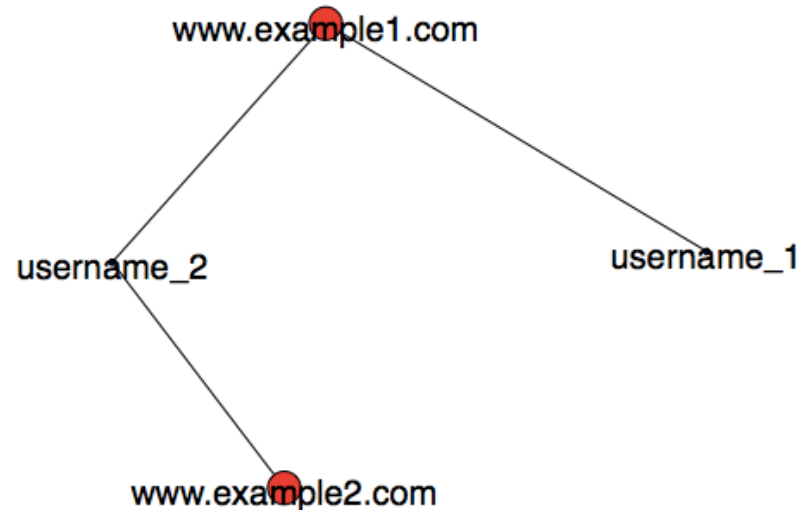
80% of PTs are active for 2 days or less

PTs that stay alive for a long time are:

- places (i.e. New York)
- Known corporations (i.e. Disney)

The Method

For each day create a graph with edges among users and links



For each user collect:

- Total Tweets
- Total, Unique PTs
- Total, Unique HT (#)
- Total, Unique UM (@)
- #followers/followings
- #URLs
- Number of Active Days

For each spam user collect:

- #Spam tweets
- #Blacklist Hits

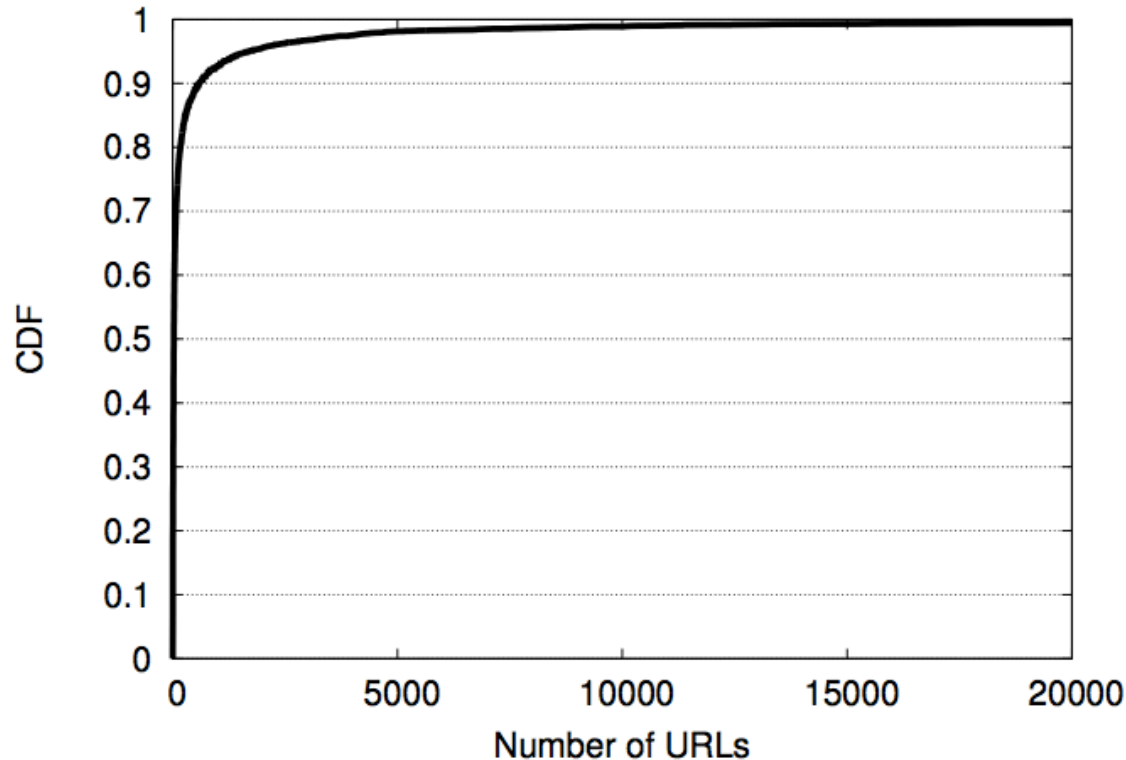
For each URL collect:

- Total, Unique PTs of the tweets that posted this URL

For each edge collect:


- #Tweets of this user that contain this URL

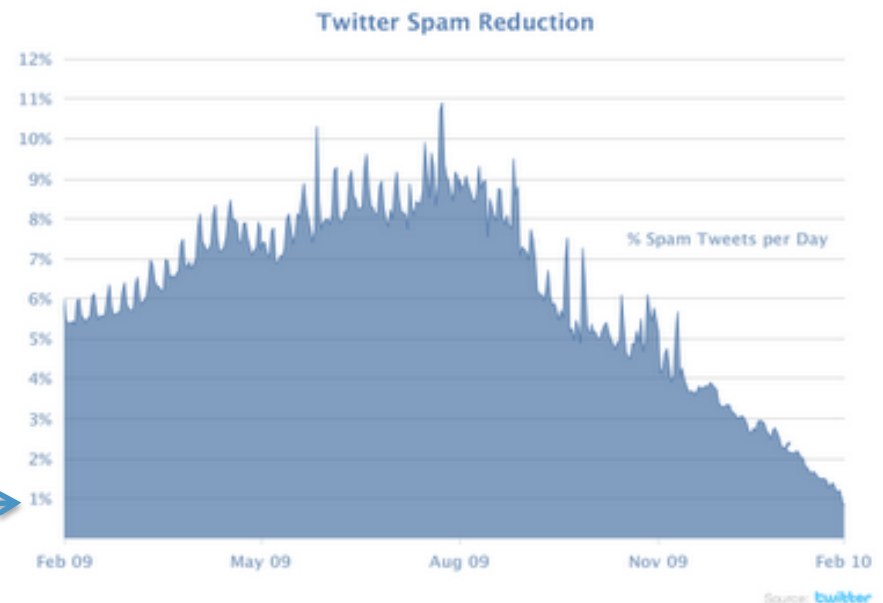
How many URLs are posted for each PT?



- 90% of PTs are associated with less than 1000 URLs
- PTs with high number of URLs: places and known corporations

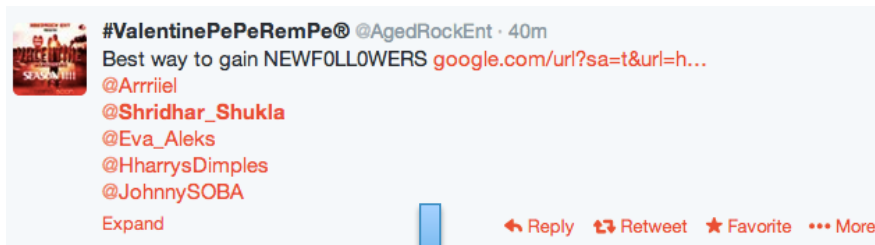
Locate spam URLs

- Get links posted by ≥ 10 users
- Extract the domains of these links
→ 24.000 domains
- Check these domains in 86 different RBLs (Real Time Blackhole Lists)
- 1.911 domains blacklisted (91 False Positives, manually checked)
- Out of 4.5M different URLs, 250K belonged to spam domain (5.4%)
 - **Twitter reports 1%..** 



Is that all spam? No there is more..

Spam URLs are hidden in a Google search results link



[www.google.com.tr/url?
sa=t&rct=j&q=&esrc=s&frm=1&source=web&cd=4&cad=rja
&sqj=2&ved=0CEAQFjAD&url=http%3A%2F
%2F**www.followmania.us**
%2F&ei=FveHUoDBHITdswa2q4HYCw&usg=AFQjCNG1f
WSqWsWxdI2QCMSeu3WGXHEtaW&sig2=fEuJc66pEvdq
9vLp2XyGdg&bvm=bv.56643336,d.Yms ...](http://www.google.com.tr/url?sa=t&rct=j&q=&esrc=s&frm=1&source=web&cd=4&cad=rja&sqj=2&ved=0CEAQFjAD&url=http%3A%2F%2Fwww.followmania.us%2F&ei=FveHUoDBHITdswa2q4HYCw&usg=AFQjCNG1fWSqWsWxdI2QCMSeu3WGXHEtaW&sig2=fEuJc66pEvdq9vLp2XyGdg&bvm=bv.56643336,d.Yms...)



Spread the love and help us grow
Tweet +1 Mou apεα

followmania.us

Signup as a Premium Member	Login as a Regular Tweeter
Now only £4.95	Always Free
New followers every minute	48 new followers per ride
Ad free and instant activation	Promotional status updates
Sign up for Premium	Sign in with twitter

Looking for free Facebook likes? Try our new service SpreadyourLikes.

Check out our latest 48 riders: (We will never send messages or alter account details)



106 Get More Followers (GMF) domains

Spam users

- Now that we know which domains are spam
- Locate spam users
 - 8.2M users → 590K spammers (*spammers & compromised users*) → 7.2%

Next step:

- discover behaviors that can differentiate between SPAM and LEGIT users

Twitter features: Spam(S) / Legit(L)



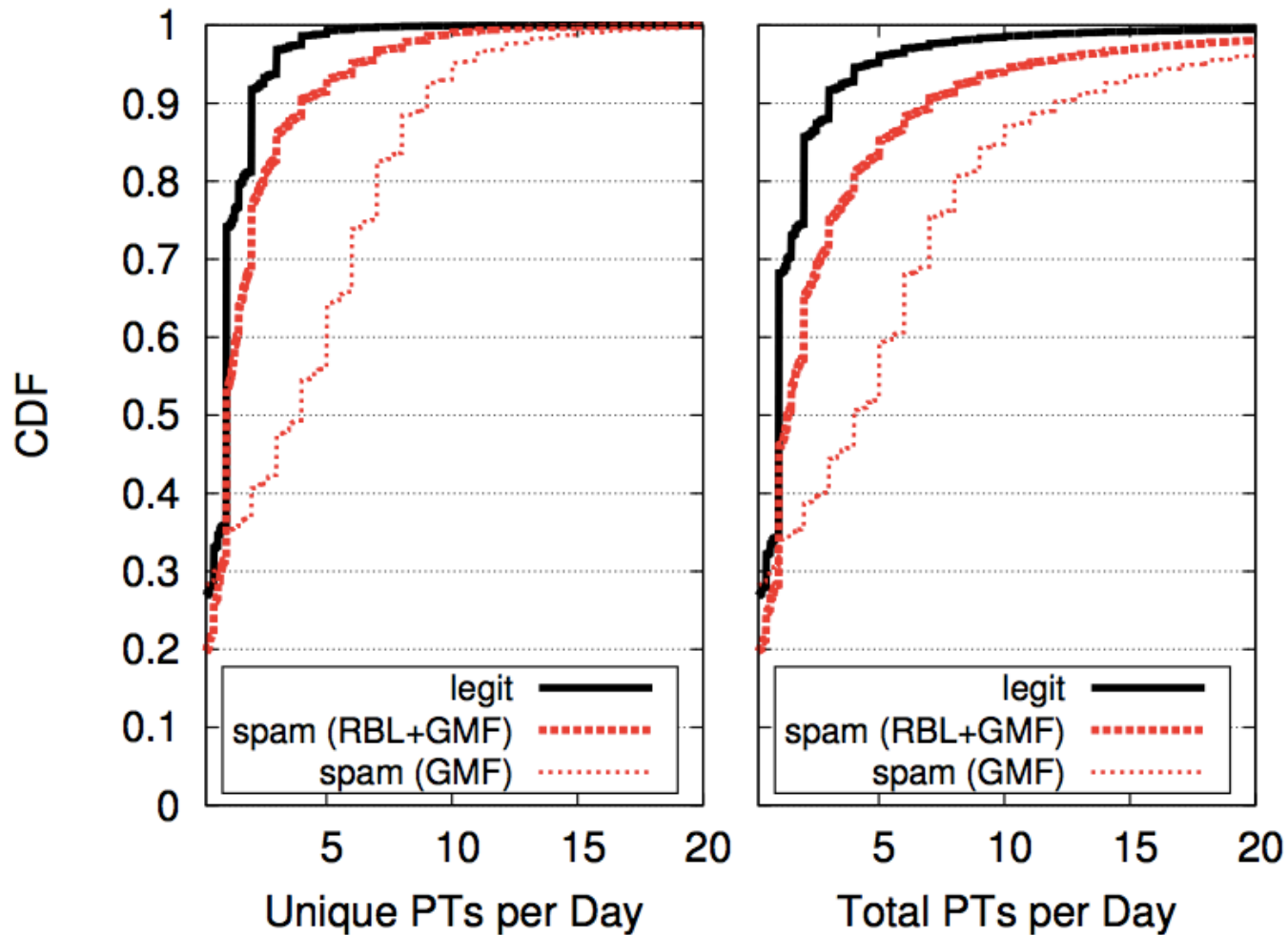
	Mean		Median	
	S	L	S	L
Tweets	2.5	1.6	1.3	1.0
Active Days	3.9	1.6	2.0	1.0
Trends (T)	2.64	1.3	1.3	1.0
Trends (U)	1.6	0.9	1.0	1.0
UM (T)	1.8	1.4	1.0	1.0
UM (U)	1.4	1.18	1.0	1.0
HashTags (T)	2.6	1.3	1.0	0.5
HashTags (U)	1.6	0.99	1.0	0.5
URLs (T)	1.6	1.1	1.0	1.0
URLs (U)	1.1	1.1	1.1	1.1
Spam URLs	1.6	NA	1.0	NA
Blacklist hits	1.9	NA	1.0	NA

Blacklists: RBL+GMF

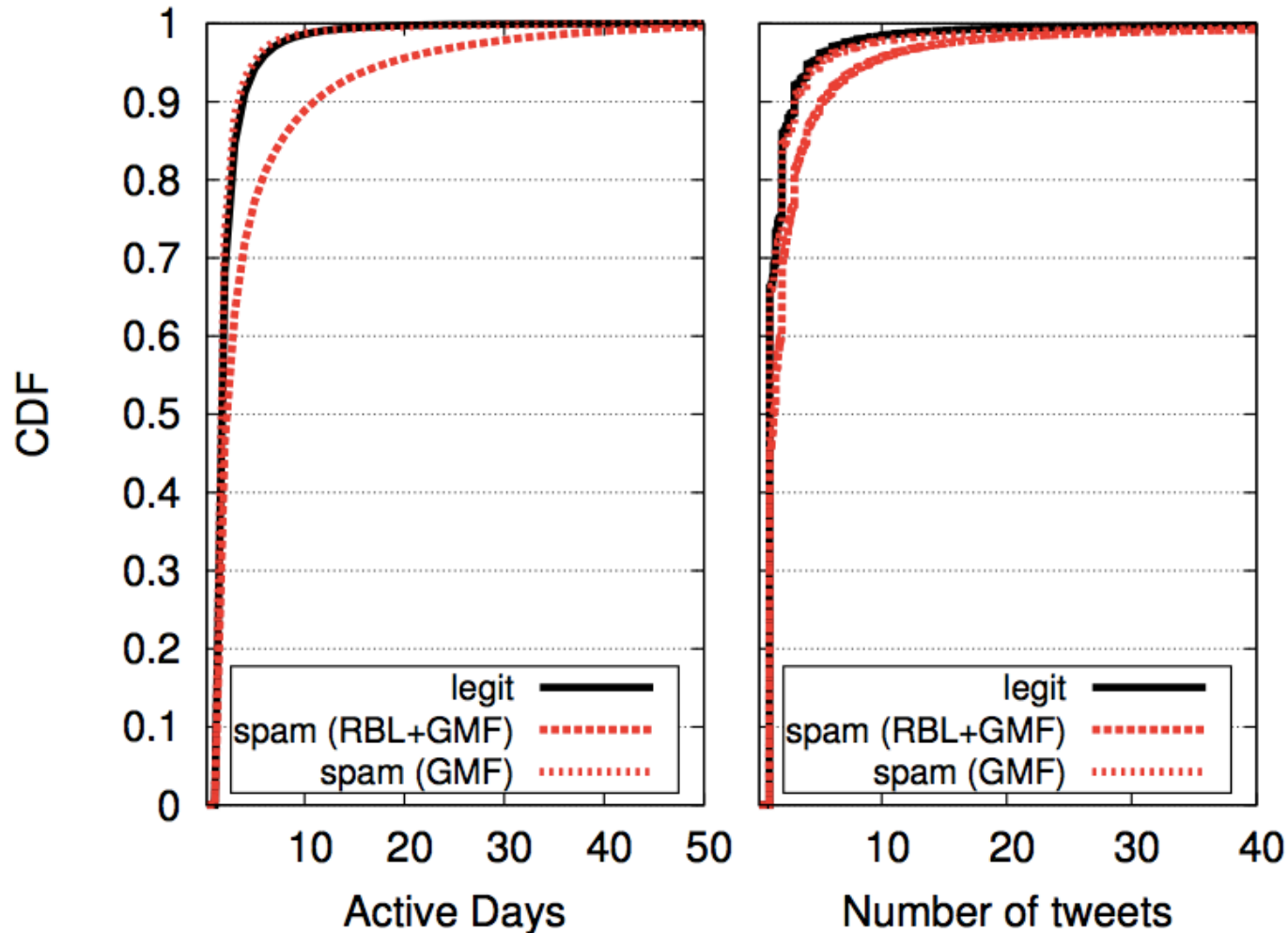
	Mean		Median	
	S	L	S	L
Tweets	1.7	1.6	1.0	1.0
Active Days	1.5	1.8	1.0	1.0
Trends (T)	5.0	1.3	4.0	1.0
Trends (U)	3.8	0.9	4.0	1.0
UM (T)	0.6	1.4	0.0	1.0
UM (U)	0.4	1.2	0.0	1.0
HashTags (T)	4.4	1.3	4.0	0.5
HashTags (U)	3.4	1.0	4.0	0.5
URLs (T)	1.4	1.1	1.0	1.0
URLs (U)	1.1	1.1	1.1	1.1
Spam URLs	1.4	NA	1.0	NA
Blacklist hits	NA	NA	NA	NA

Blacklists: GMF

Spammers exploit PTs



RBL spammers are more active than GMF (this is because GMF spammers are people too..)



Classification

- Classify users based on Twitter features
- Train: February 2014, Test: March 2014
- Method: Decision Trees

$$\textit{Sensitivity} = \frac{TP}{TP + FN}$$

$$\textit{Specificity} = \frac{TN}{FP + TN}$$

Sensitivity: spammers correctly classified,
Low means that users are not adequately
protected by spam...

Still better User Experience than before

Specificity: legitimate classified correctly,
Low means legit users incorrectly as
spammers...

Worse User Experience than before

Classification Results

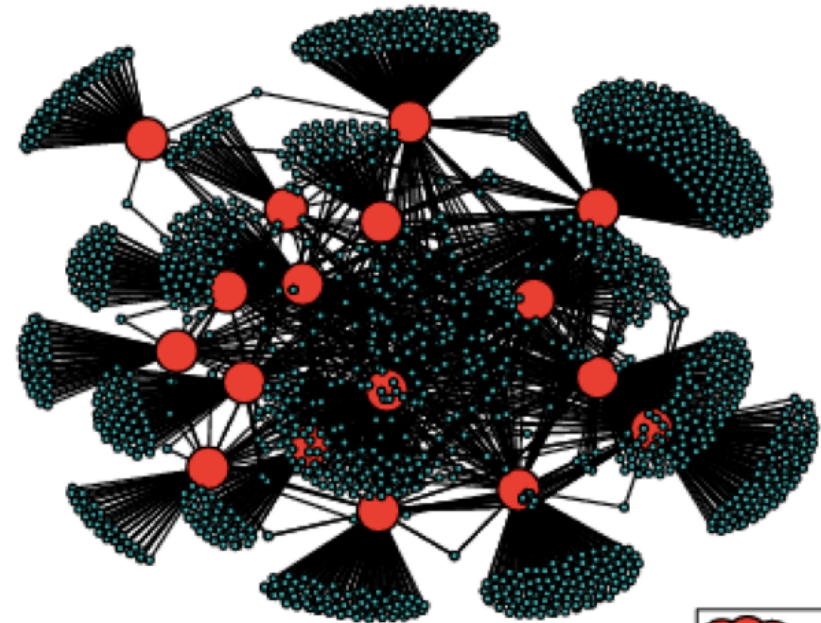
Objective	Blacklist	Sensitivity	Specificity
Users based on Twitter features	RBL + GMF	13,3% 40,808/306,005	98.12% 4,373,684/4,457,565
	GMF	37,6% 19,674/52,281	99.92% 4,707,627/4,711,289
URLs based on PTs	GMF	45.6% 317/696	99.83% 48353/48435

- Completely independent Test and Train samples (no cross validation)
- Very simple computation. Easy to apply.
- Due to high Specificity, can be a first line of defense

Spam campaigns

- January 10th 2014
- 2 IPs(!) → 17 GMF domains → 1604 users

Domain	IP Addr.	Users	Tweets
bestfollowers.org	216.55.x.y	337	462
goodfollowers.net	216.55.x.y	188	223
getmorefollowers.biz	216.55.x.y	179	283
newfollowers.me	216.55.x.y	175	207
justfollowers.us	216.55.x.y	152	173
getnewfollowers.us	216.55.x.y	151	195
followcrazy.us	216.55.x.y	150	161
bulkfollowers.co	216.55.x.y	131	169
followport.us	216.55.x.y	99	121
followcity.us	216.55.x.y	93	109
morefollowers.me	216.55.x.y	81	96
twitterfollowers.mobi	216.55.x.y	76	98
t1t.us	68.178.x.y	74	108
followmania.us	216.55.x.y	68	86
specialfollowers.com	216.55.x.y	64	73
livefollowers.org	216.55.x.y	67	77
onlinefollowers.com	216.55.x.y	40	51



URL node average degree:
Spam: 125
Legit: 2.3



Discussion

- Framework to model and store a rich set of Twitter features
- Popular Trends is a feature widely exploited by spammers
- Blacklists are still useful although link obfuscation techniques can bypass them (like Google results link)
- 5% of URLs and 7% of users in Twitter are spam
- Dedicated spam accounts and exploited real accounts have different tweeting behaviors
- Twitter features can be used to build a computational efficient first line of defense (More than 1/3 TPR, 99.9% TNR)
- Half of spam URLs can be identified by using only PT information
- Graph properties of Spam campaigns is a powerful method to detect and study spam in Twitter.

Submitted ...



Future work:

Decentralized Spam Filtering

Adjust the procedure in a gossip algorithm:

- Trend and Tweet collection is already distributed among multiple accounts
- Store in MongoDB
- Run ensemble methods for classification:
 - Adaboost
 - Random Forest Classifier
- Merge campaign graphs in order to identify larger spam campaigns,
- “get the complete picture”