Social Spam, Campaigns, Misinformation and Crowdturfing

Kyumin Lee
Utah State Univ.

James Caverlee
Texas A&M Univ.

Calton Pu
Georgia Tech

April 7, 2014 @ WWW 2014

Schedule

14:00 ~ 14:10	Introduction to Social Media Threats (Social Spam, Campaigns, Misinformation and Crowdturfing)
14:10 ~ 14:55	Social Spam
14:55 ~ 15:30	Campaigns
15:30 ~ 16:00	30 min Break
16:00 ~ 16:30	Misinformation
16:30 ~ 17:10	Crowdturfing
17:10 ~ 17:30	Challenges, Opportunities and Conclusion

Disclaimers

- Since the tutorial is only 3 hours long, we will focus on presenting social media threats and countermeasures of recent research results.
- But, we don’t have time to give great depth on every possible result, so we will highlight a few representatives.
- We will provide many relevant references in the end of the tutorial.

Schedule

14:00 ~ 14:10	Introduction to Social Media Threats (Social Spam, Campaigns, Misinformation and Crowdturfing)
14:10 ~ 14:55	Social Spam
14:55 ~ 15:30	Campaigns
15:30 ~ 16:00	Break
16:00 ~ 16:30	Misinformation
16:30 ~ 17:10	Crowdturfing
17:10 ~ 17:30	Challenges, Tools and Conclusion

Large-Scale Social Systems

Online Social Networking	facebook	twitter	LinkedIn	
Social Media	YouTube	flickr	digg	
Information sharing communities	reddit	YAHOO! ANSWERS	StumbleUpon	
Social Games	zynga	ROVIO	EA	wooga world of gaming
Location-based Services	foursquare	yelp	Google Latitude mobile / iGoogle	Gowalla
Crowd-based services	CrowdFlower	KICKSTARTER	INDIEGOGO Where Independent Happens	

Large-Scale Social Systems: **Key Organizing Principles**

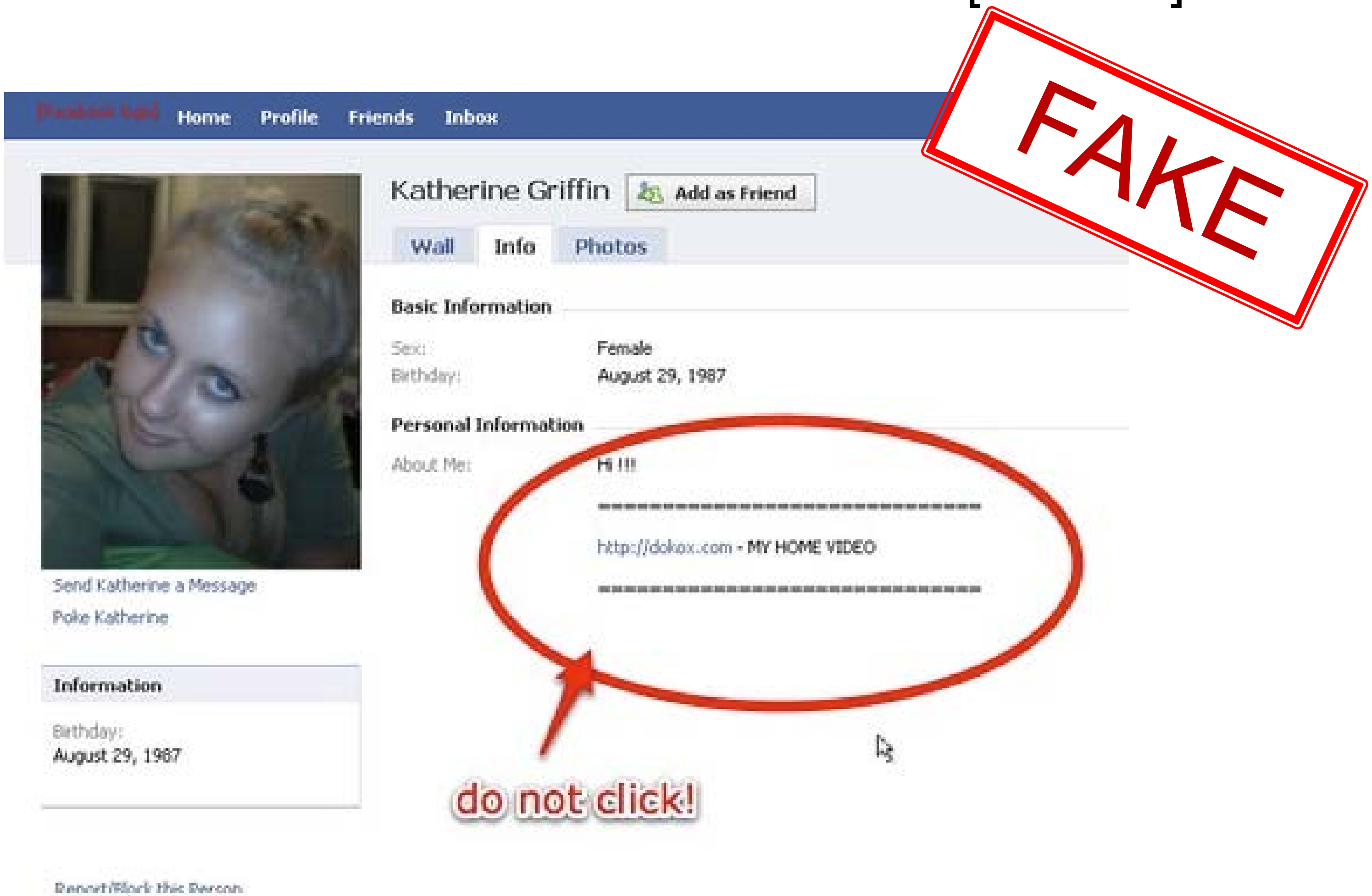
- Openness:
  - Social systems are inherently open to users who generate, share and consume information
  - E.g., post a message, upload and watch a video
- Collaboration:
  - Many users organically participate in social systems to engage in collaborative activities
  - E.g., organize for political change, share disaster-related information
- Real-time information propagation:
  - Users, media and organization post information related to hot events in (near) real-time
  - E.g., emergency alerts, natural disaster news and sports games
- Crowdsourcing tasks or hiring cheap workers from all over the world:
  - People can hire workers from crowdsourcing sites with paying little money
  - E.g., workers from Amazon Mechanical Turk for labeling data, workers from Fiverr for editing a document

Large-Scale Social Systems: **Challenges and Research Approach**

- These necessary positive aspects may also lead to negative consequences
  - Spam of many flavors
    - Comment spam (~90% on websites = 46 billion)
    - Spam tweets (1% = 3 million/day) and Twitter spammers (5% = 25 million)
    - Spam videos (20%)
  - Traditional Attacks
    - Phishing, malware and etc
  - Campaigns
  - Misinformation
  - Crowdturfing
  - Misuse
    - Crowdsourcing the wrong guy in the Boston bombings at Reddit
  - …

Fake Accounts

- 9% on Facebook = 87 million accounts in 2012 [Facebook]



Comment Spam

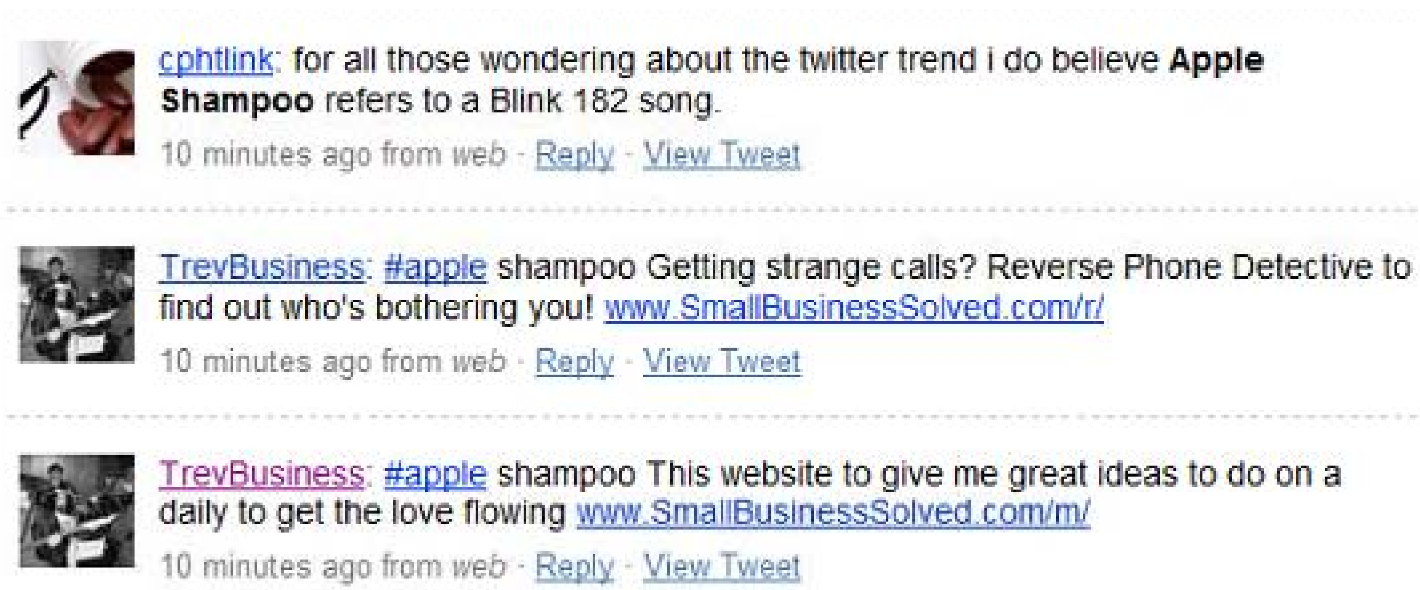
- 83 ~ 90% on websites = 46 billion comments [Akismet and Mollom. 2010, Kant et al. WSDM 2012]

Rosiane facebook.com/profile.php?id=100003406202721 x m.smealen@mail.ru 188.143.232.12	Submitted on 2012/07/02 at 09:27 you people may not belivee at all but i can and will tell you that between heaven and earth are things beyond the reach of ordinary man and women.you people do not know what knowledge is and you would not gain any knowledge if its not by some devine revelation.is this the book of the devil maybe but it sure as hell is not for ordinary folks like you people to read, you could not handle it any one of you, before you open the book of the devil you better make sure your in a right pad with GOD Jehova.
Urvi facebook.com/profile.php?id=100003406194827 x info@sms-vluchtelingen.nl 188.143.232.12	Submitted on 2012/07/02 at 02:20 I had a spambot at my potrey site post something regarding the size of her husband.All I can say is Mr. Jeremy must be glad he isn't married to her.Then there's the one with the guy wanting to sell his bridal dresses.
best affiliate website home-businessreviews.com/Turnkey-Affiliate-Websit... x justinjki111558@gmail.com 46.109.196.107	Submitted on 2012/06/29 at 04:34 Make $1,000's Weekly with a Health Internet Business of Your Very Own Now get a complete fully-operational "Health eBiz" in a box! This amazing site: * Closes sales automatically for you! * Has a complete electronic sales manager that makes all upsells for you!

Spam Tweets and Twitter Spammers

- 1% Spam tweets and 5% Twitter spammers
  - 3 million spam tweets/day and 25 million spam accounts

    [Twitter and TwitSweeper, 2010]

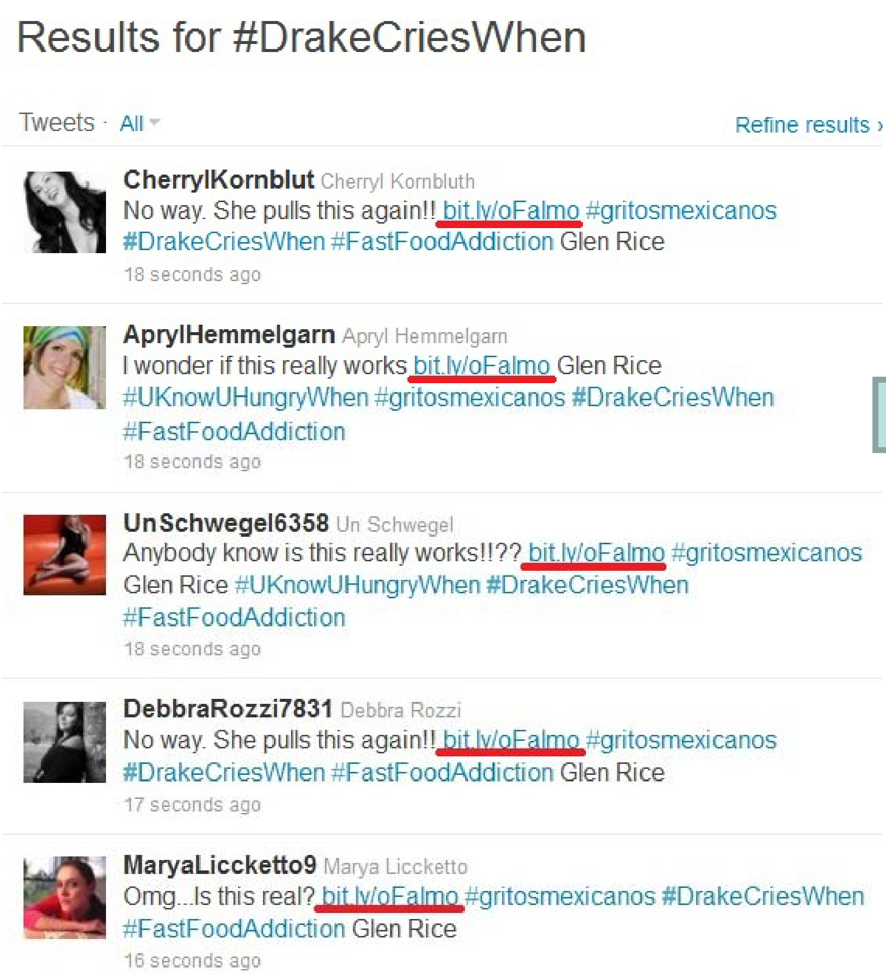


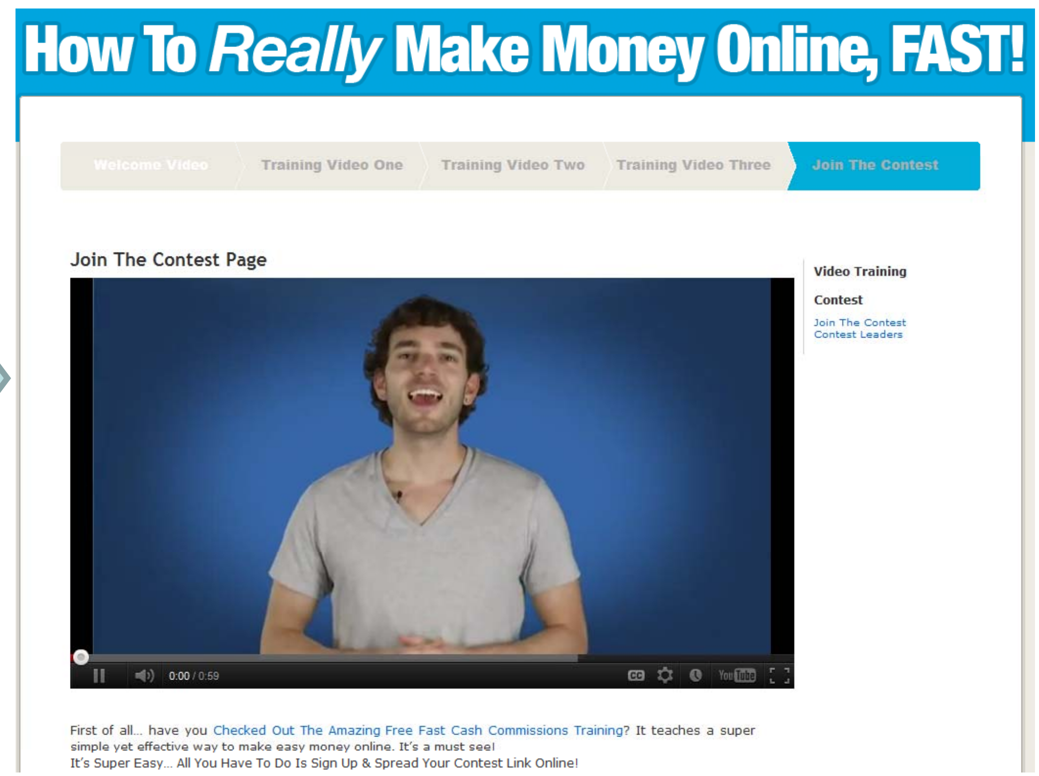
Spam Videos

- 183 million U.S. Internet users watched more than 37 billion online videos in Oct 2012. [comScore]
- 20% of online videos are spam [VideoSurf]

Collective Attention Spam

- Target popular and trendy topics/items
- Feed spam contents once the topics/items become popular





Campaigns

Astroturfing

The need to protect the internet from 'astroturfing' grows ever more urgent

The tobacco industry does it, the US Air Force clearly wants to … astroturfing – the use of sophisticated software to drown out real people on web forums – is on the rise. How do we stop it?

A real person using the internet. Unfortunately we can no longer assume what we are reading is written by one of these creatures. Photograph: Jeff Blackler/Rex Features

Fake review campaign

Big John's Profile	Cletus' Profile	Jake's Profile
1 of 1 people found the following review helpful: ★★★★★ Practically FREE music, December 4, 2004 This review is from: Audio Xtract (CD-ROM) I can't believe for $10 (after rebate) I got a program that gets me free unlimited music. I was hoping it did half what was ….	2 of 2 people found the following review helpful: ★★★★★ Like a tape recorder…, December 8, 2004 This review is from: Audio Xtract (CD-ROM) This software really rocks. I can set the program to record music all day long and just let it go. I come home and my ….	★★★★★ Wow, internet music! …, December 4, 2004 This review is from: Audio Xtract (CD-ROM) I looked forever for a way to record internet music. My way took a long time and many steps (frustrtaing). Then I found Audio Xtract. With more than 3,000 songs downloaded in …
3 of 8 people found the following review helpful: ★★★★★ Yes – it really works, December 4, 2004 This review is from: Audio Xtract Pro (CD-ROM) See my review for Audio Xtract - this PRO is even better. This is the solution I've been looking for. After buying iTunes, ….	3 of 10 people found the following review helpful: ★★★★★ This is even better than…, December 8, 2004 This review is from: Audio Xtract Pro (CD-ROM) Let me tell you, this has to be one of the coolest products ever on the market. Record 8 internet radio stations at once, ….	2 of 9 people found the following review helpful: ★★★★★ Best music just got …, December 4, 2004 This review is from: Audio Xtract Pro (CD-ROM) The other day I upgraded to this TOP NOTCH product. Everyone who loves music needs to get it from Internet ….
5 of 5 people found the following review helpful: ★★★★★ My kids love it, December 4, 2004 This review is from: Pond Aquarium 3D Deluxe Edition This was a bargain at $20 - better than the other ones that have no above water scenes. My kids get a kick out of the ….	5 of 5 people found the following review helpful: ★★★★★ For the price you…, December 8, 2004 This review is from: Pond Aquarium 3D Deluxe Edition This is one of the coolest screensavers I have ever seen, the fish move realistically, the environments look real, and the ….	3 of 3 people found the following review helpful: ★★★★★ Cool, looks great…, December 4, 2004 This review is from: Pond Aquarium 3D Deluxe Edition We have this set up on the PC at home and it looks GREAT. The fish and the scenes are really neat. Friends and family ….

Political campaign

COMPUTING

Bogus Grass-Roots Politics on Twitter

Data-mining techniques reveal fake Twitter accounts that give the impression of a vast political movement.

TUESDAY, NOVEMBER 2, 2010 | BY KURT KLEINER

Audio »

How true? This network graph shows the connections between 6,278 accounts that used the hashtag #gop in September and October 2010.
Indiana University

Researchers have found evidence that political campaigns and special-interest groups are using scores of fake Twitter accounts to create the impression of broad grass-roots political expression. A team at Indiana University used data-mining and network-analysis techniques to detect the activity.

"We think this technique must be common," says Filippo Menczer, an associate professor at Indiana University and one of the principal investigators on the project. "Wherever there are lots of eyes looking at screens, spammers will be there; so why not with politics?"

Website	Campaigns	% Crowdturfing	Tasks	$ per Subm.
Amazon Turk (US)	41K	12%	2.9M	$0.092
ShortTask* (US)	30K	95%	527K	$0.096
MinuteWorkers (US)	710	70%	10K	$0.241
MyEasyTask (US)	166	83%	4K	$0.149
Microworkers (US)	267	89%	84K	$0.175

Wang et al. WWW 2012

Adversarial Propaganda

- Create and spread rumors and Misinformation
- Target a product/ government

Pentagon Wants a Social Media Propaganda Machine

BY ADAM RAWNSLEY 07.15.11 2:40 PM

Follow @arawnsley

Like 1.4k
Tweet 835
+1 113
Share 165

You don't need to have 5,000 friends of Facebook to know that social media can have a notorious mix of rumor, gossip and just plain disinformation. The Pentagon is looking to build a tool to sniff out social media propaganda campaigns and spit some counter-spin right back at it.

On Thursday, Defense Department extreme technology arm Darpa unveiled its Social Media in Strategic Communication (SMISC) program. It's an attempt to get better at both detecting and

[Wired]

Misinformation (Fake)

DC Maryland Virginia
@DMVFollowers
Follow

McDonalds in Virginia Beach flooded. pic.twitter.com/FZBoCydM

Reply Retweet Favorite

Fake Images

Katina
@kdekranis9

I TOLD Y'ALL! Shark on the highway in New Jer[...] @maxthewanted would appreciate this. #Hurric[...] pic.twitter.com/kaYMjWzT

1:09 AM - 30 Oct 2012

Jamster
@jamster83
Follow

Amazing picture of hurricane #Sandy decending in New York pic.twitter.com/3mMhCbNq

4:21 PM - 29 Oct 2012

2,745 RETWEETS 586 FAVORITES

Crowdturfing (Crowdsourcing + Astroturfing)

- A Multimillion-dollar industry in Chinese crowdsourcing sites
  - 90% crowdturfing tasks [MIT Technology Review]
- 70~95% crowdturfing tasks at several U.S. crowdsourcing sites [Wang et al., WWW 2012]

Twitter Post: CPP Scam

Work done: **222/250**
Employer: Member_968289

You will earn **$0.60**

This task takes less than **30** min to finish

Job ID: 364488d297e8

? What is expected from Workers?

You must have 50 Twitter followers. Make sure you are logged into your Twitter account

1. Open your browser and search on Google "college pro painters success"
2. Click on any search result that starts with c o l l e g e p r o p a i n t e r s s c a m . c o m

3. Go to Home Page of the website
4. Retweet any article

Examples of Crowdturfing

- Vietnamese propaganda spread by 1,000 crowdturfers

Vietnam admits deploying bloggers to support government

By Nga Pham
BBC News, Hanoi

Vietnamese propaganda officials have admitted deploying people to engage in online discussions and post comments supporting the Communist Party's policies.

The party has also confirmed that it operates a network of nearly 1,000 "public opinion shapers".

They are assigned with the task of spreading the party line.

The tactic is similar to China's model of internet moderators who aim to control news and manipulate opinion.

'Political opportunists'

Hanoi Propaganda and Education Department head Ho Quang Loi said that the authorities had hired hundreds of so-called "internet polemists" in the fight against "online hostile forces".

The bloggers have been hailed for stopping negative online rumours

Examples of Crowdturfing

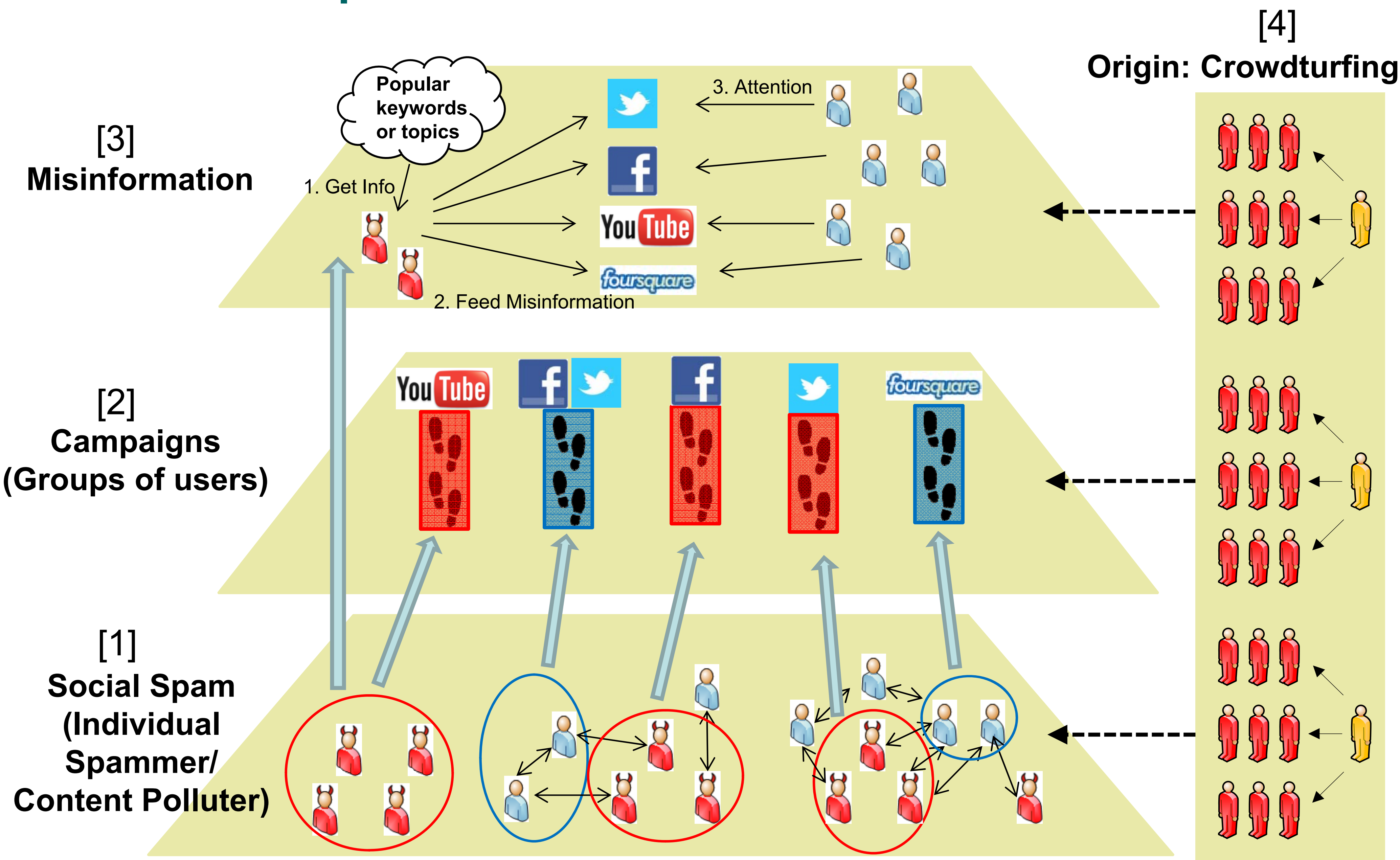
CHINADAILY

"Dairy giant Mengniu in smear scandal"

Warning: Company Y's baby formula contains dangerous hormones!

- Biggest dairy company in China (Mengniu)
  - Defame its competitors
  - Hire Internet users to spread false stories
- Impact
  - Victim company (Shengyuan)
    - Stock fell by 35.44%
    - Revenue loss: $300 million
  - National panic

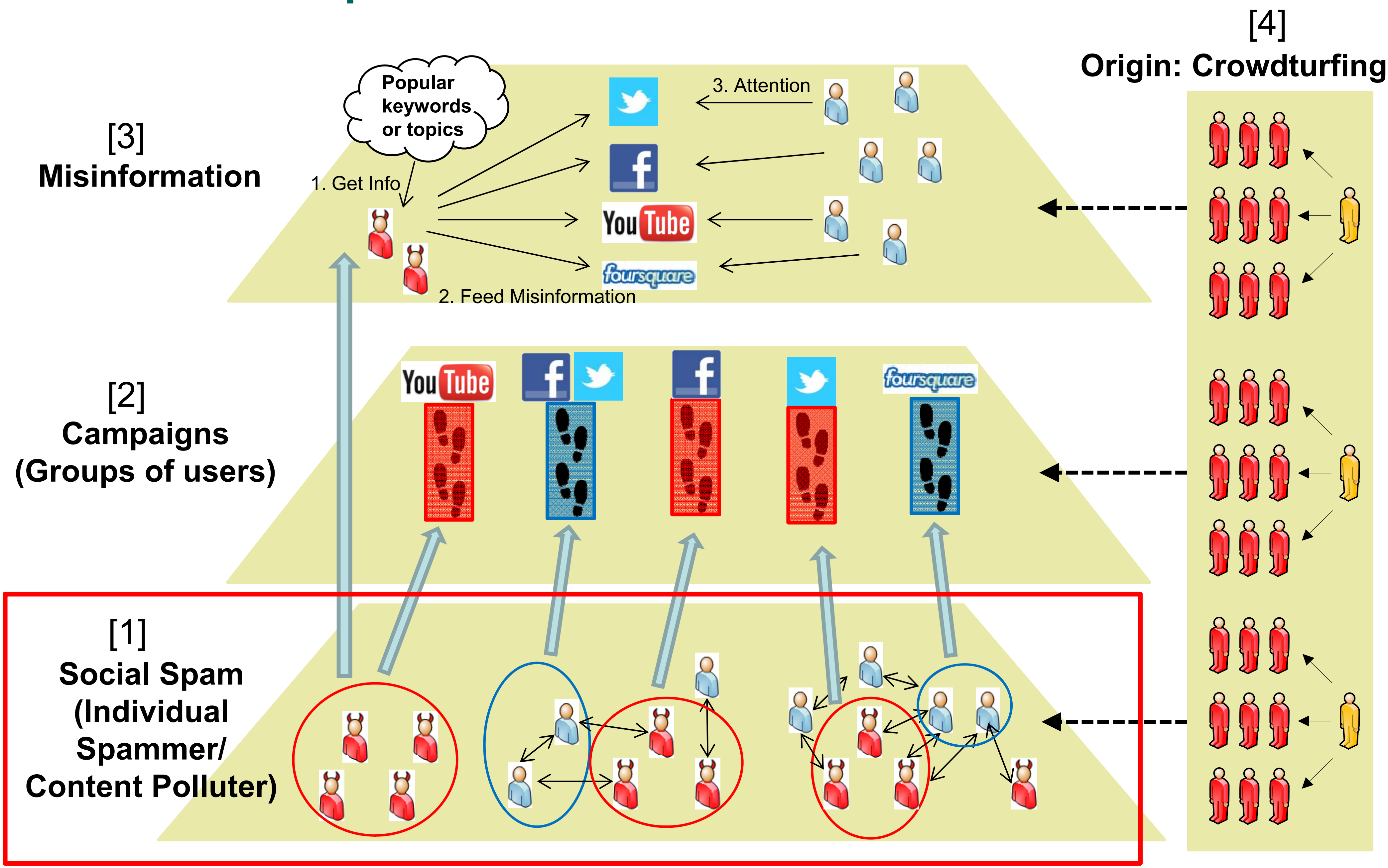
Conceptual Level of Tutorial Theme



Schedule

14:00 ~ 14:10	Introduction to Social Media Threats (Social Spam, Campaigns, Misinformation and Crowdturfing)
14:10 ~ 14:55	Social Spam
14:55 ~ 15:30	Campaigns
15:30 ~ 16:00	Break
16:00 ~ 16:30	Misinformation
16:30 ~ 17:10	Crowdturfing
17:10 ~ 17:30	Challenges, Opportunities and Tools in Social Spam, Campaigns, Misinformation and Crowdturfing Research

Conceptual Level of Tutorial Theme



Social Spam

- Fake accounts (5 ~ 6 % on Facebook = 42 million)
  - [Facebook. 2012]
- Comment spam (83 ~ 90% on websites = 46 billion)
  - [Akismet and Mollom. 2010, Kant et al. WSDM 2012]
- Spam Tweets (1% = 3 million/day) and Twitter Spammers (5% = 25 million)
  - [Twitter. 2010, TwitSweeper. 2010, Lee et al. SIGIR 2010, Lee et. al ICWSM 2011, Yang et al. WWW 2012]
- Tag spam
  - [Koutrika et al. TWEB 2008, Krause et al. AIRWEB 2008 , Neubauer et al. AIRWEB 2009]
- Spam videos
  - [Benevenuto et al. AIRWeb 2008, Benevenuto et al. SIGIR 2009]
- Fake Reviews
  - [Jindal and Bing ICDM 2007, Lim et al. CIKM 2010, Wang et al. TIST 2011, Mukherjee et al. WWW 2012]
- Voting spam
  - [Bian et al. AIRWEB 2008, Tran et al. NSDI 2009]
- Wikipedia vandalism
  - [Potthast et al. ECIR 2008, Chin et al. WICOM 2010, Adler et al. CICLing 2011]
- …

Blacklisting URLs

- Crawled URLs from Twitter
  - 25 million URLs crawled
  - 8% of them link to spam pages

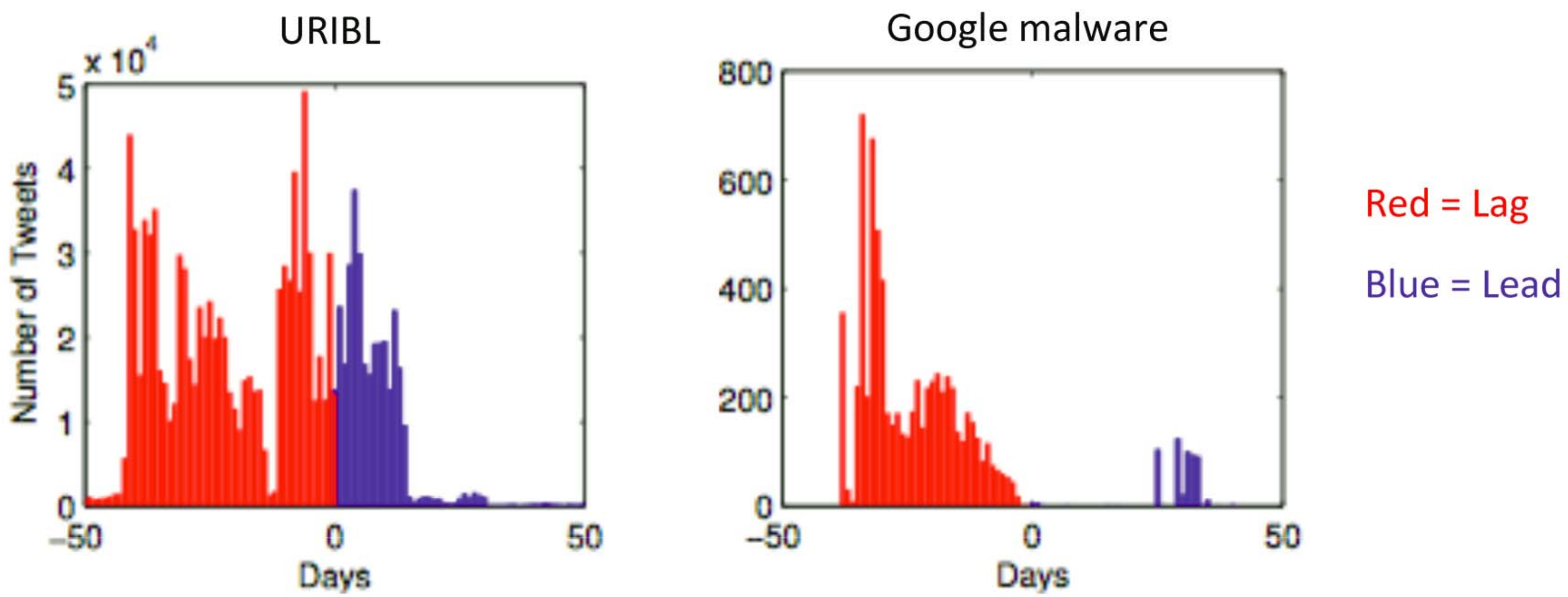
- Over 80% of spam URLs were shortened
  - Mask landing site
    - http://bit.ly/aLEmck -> http://i-drugspedia.com/pill/Viagra…
  - Defeat blacklist filtering
    - bit.ly -> short.to -> malware landing page

Grier, C., Thomas, K., Paxson, V., and Zhang, M. @spam: the underground on 140 characters or less. In CCS, 2010.

Blacklist Performance

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

- Retroactively blacklist



Comparison to Email Clickthrough

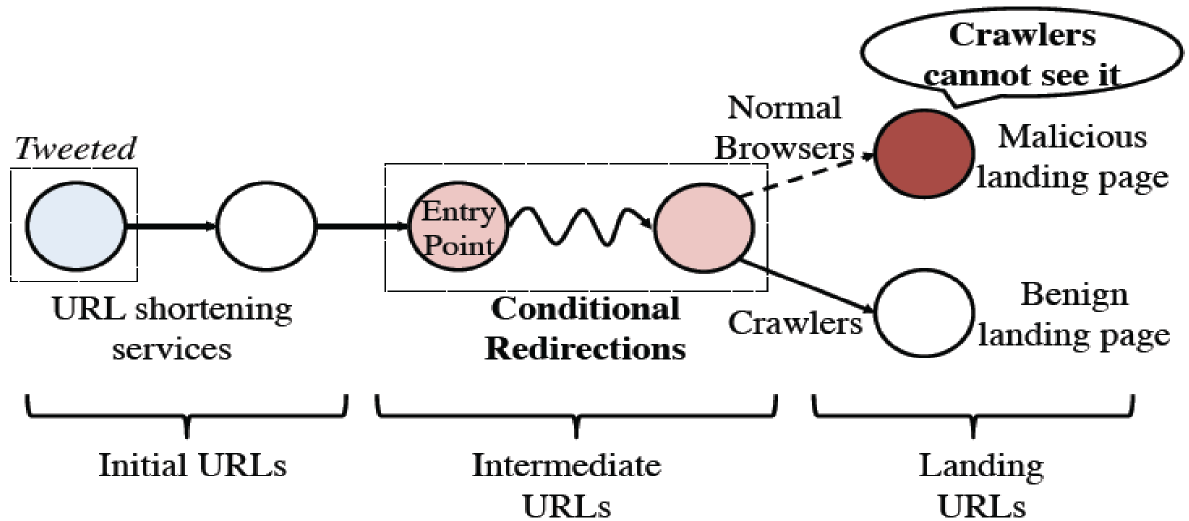
- Spam Email clickthrough: .003-.006%
  - From Spamalytics, Kanich et al. CCS 2008
- Twitter clickthrough: .13%
  - Collected 245,000 spam URLs
  - Define clickthrough as clicks / reach
  - Reach defined as *tweets* * *followers*

Social Spam Detection Approaches

- Supervised spam detection approach
  - The most popular approach
  - Require labeled data for training purpose
- Ranking users based on their social graph
- Use crowd wisdom (humans) to identify fake accounts

Supervised spam detection approach

Conditional Redirection

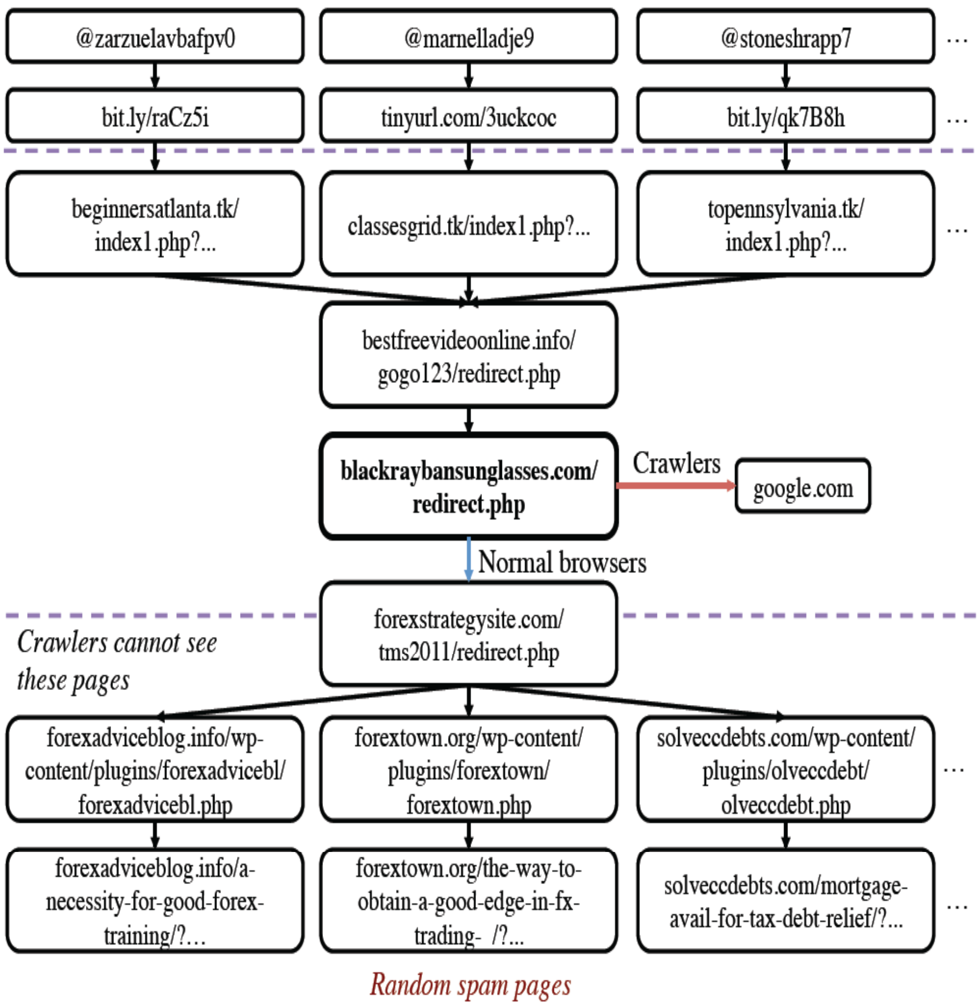


- Attackers distribute initial URLs of conditional redirect chains via tweets.
  – Initial URLs are shortened.
- Conditional redirect server will lead
  – normal browsers to malicious landing pages
  – **crawlers to benign landing pages**

Misclassifications can occur.

Lee, S., and Kim, J. WarningBird: Detecting suspicious URLs in Twitter stream. In *NDSS*, 2012

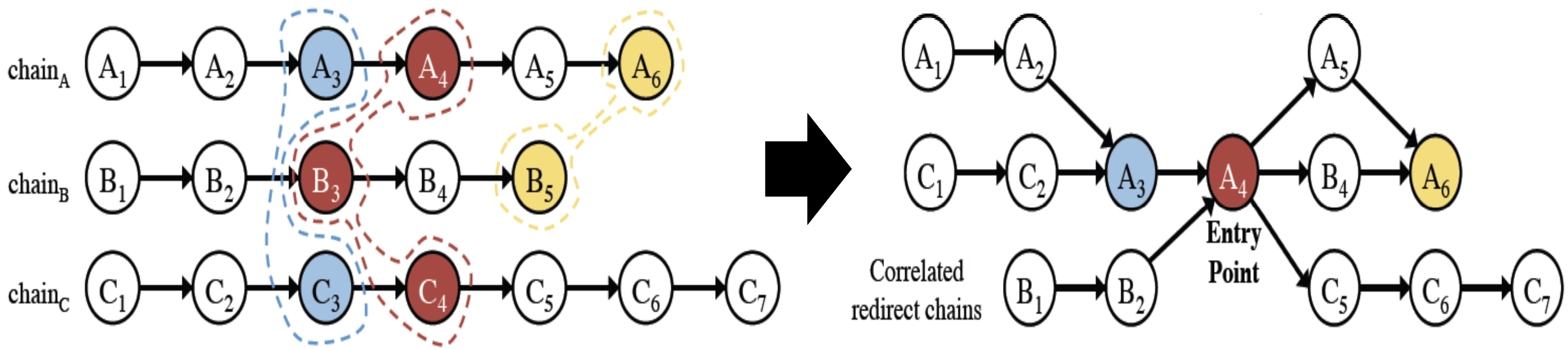
blackraybansunglasses.com



July 11, 2011

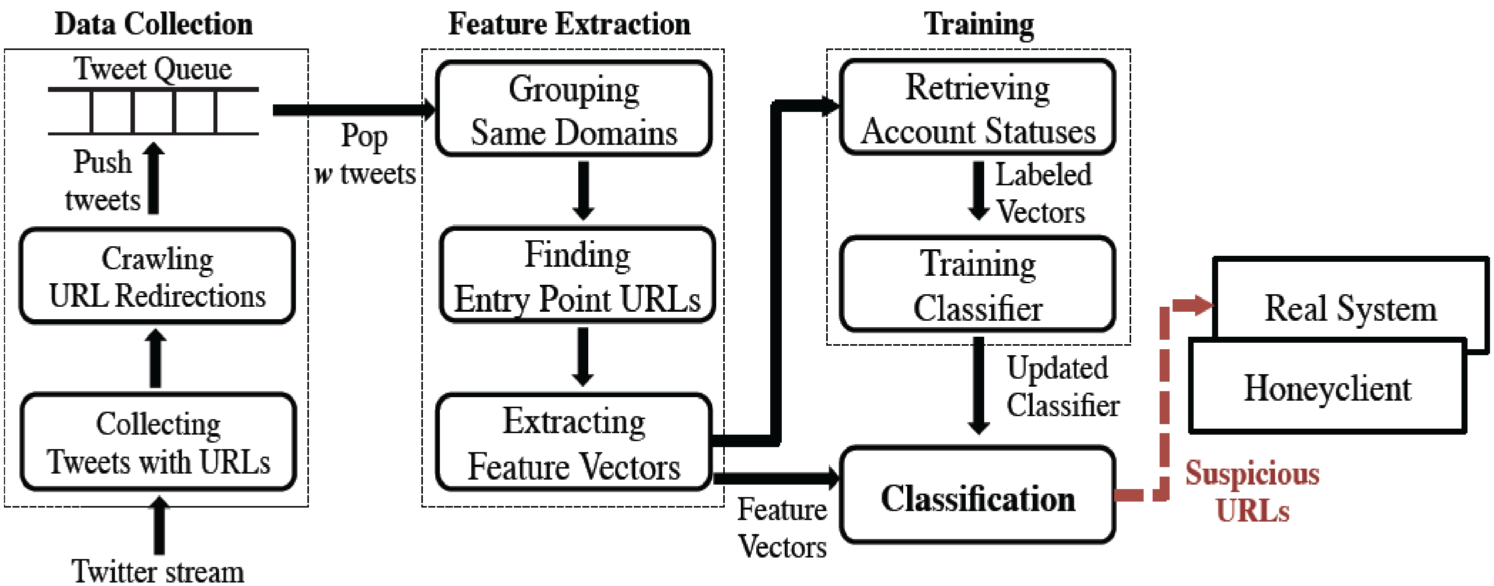
- 6,585 different accounts and shortened URLs
  - about 3% of all the daily tweets sampled
- Condition redirection
  - google.com for crawlers
  - random spam pages for normal browsers
- Some servers **reused**

Basic Idea



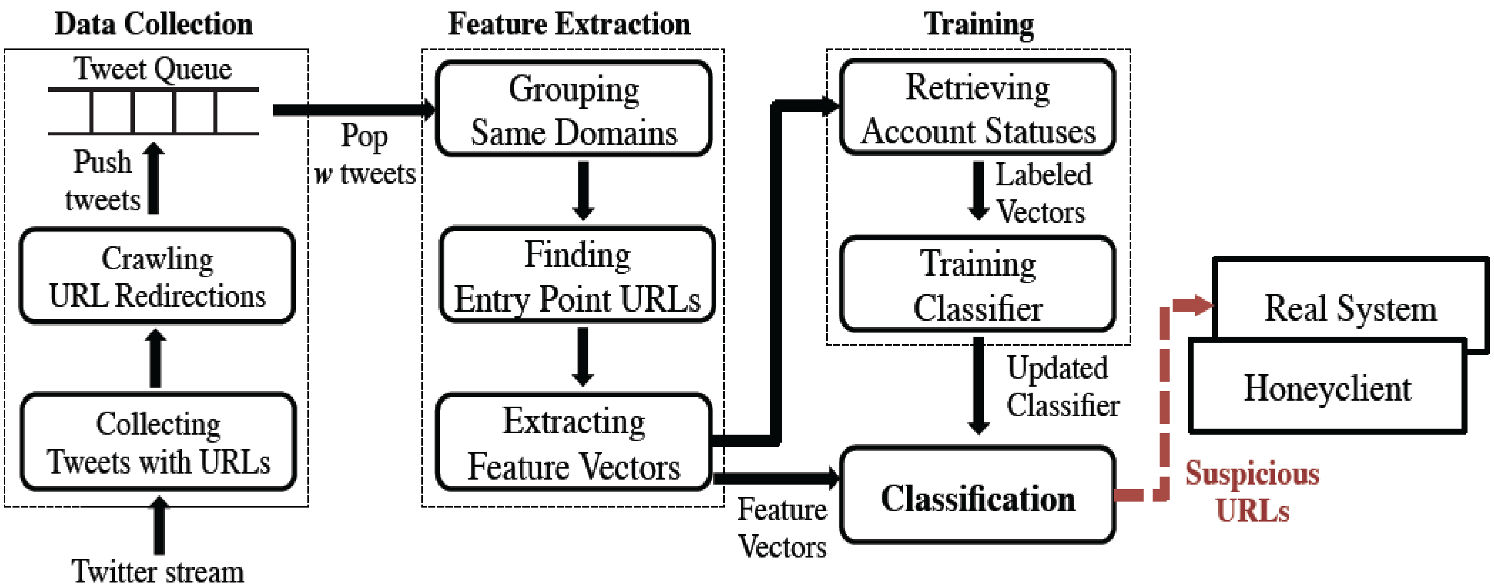
- Attackers need to **reuse** redirection servers.
  - no infinite redirection servers
- They analyze a group of correlated URL chains.
  - to detect redirection servers reused
  - to figure out features of the correlated URL chains

System Overview



- Data collection
  - collect tweets with URLs from Twitter public timeline
  - visit each URL to obtain URL chains and IP addresses
- Feature extraction
  - group domains with the same IP addresses from 10,000 tweets containing URLs
  - find entry point URLs
  - generate feature vectors for each entry point

System Overview



- Training
  - label feature vectors using account status info.
    - suspended ⇒ malicious, active ⇒ benign
  - build classification models
- Classification
  - classify suspicious URLs

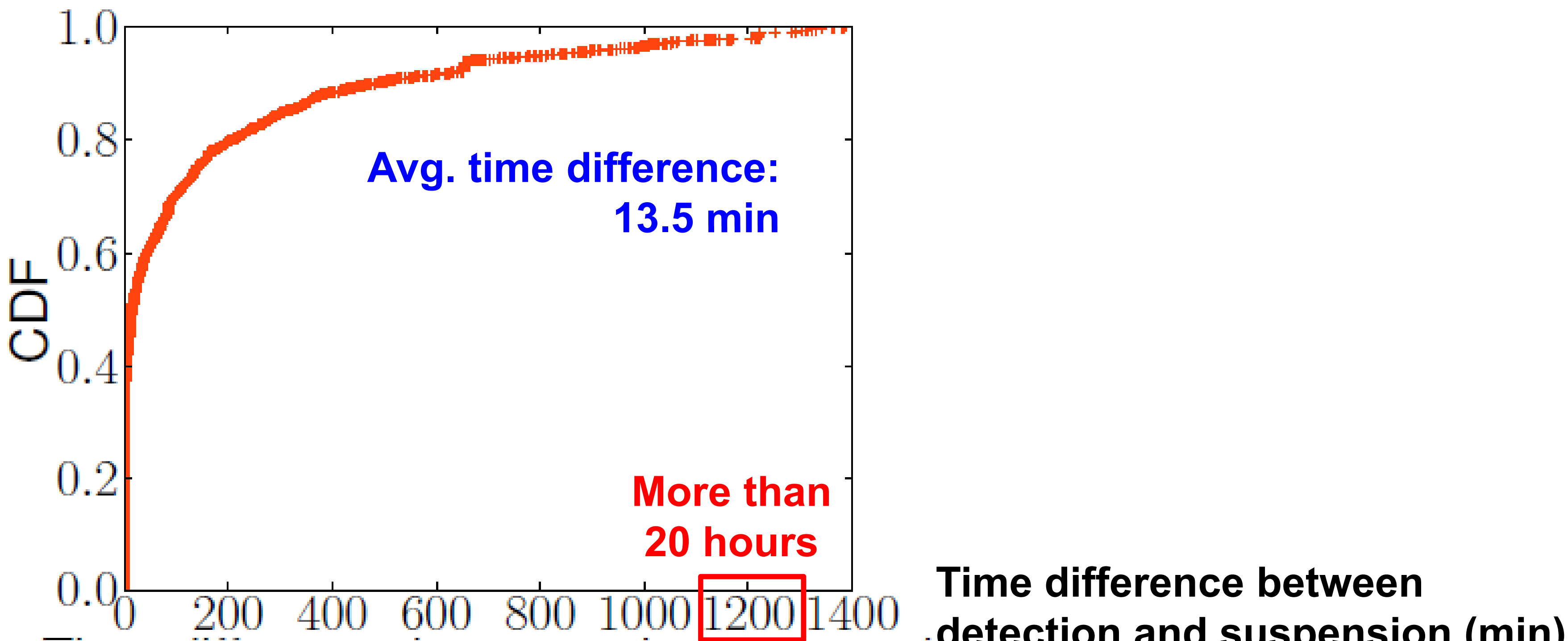
Features

- Suspiciousness of correlated URL chains
  - length of URL redirect chain
  - frequency of entry point URL
  - # of different initial and landing URLs
- Similarity of accounts posting the same URL chains
  - # of Twitter applications and accounts
  - account creation dates
  - followers-friends ratios
  - # of followers and friends

Training Classifiers

- Training dataset
  - Tweets between Sept 2011 and Oct 2011
  - 156,896 benign and 26,950 malicious entry point URLs
- Classification algorithm
  - support vector classification
  - 10-fold cross validation
  - false positive: **1.13%**, False negative: **7.01%**

Detection Efficiency



- They measure the time difference between
  - when WarningBird detects suspicious accounts
  - when Twitter suspends the accounts

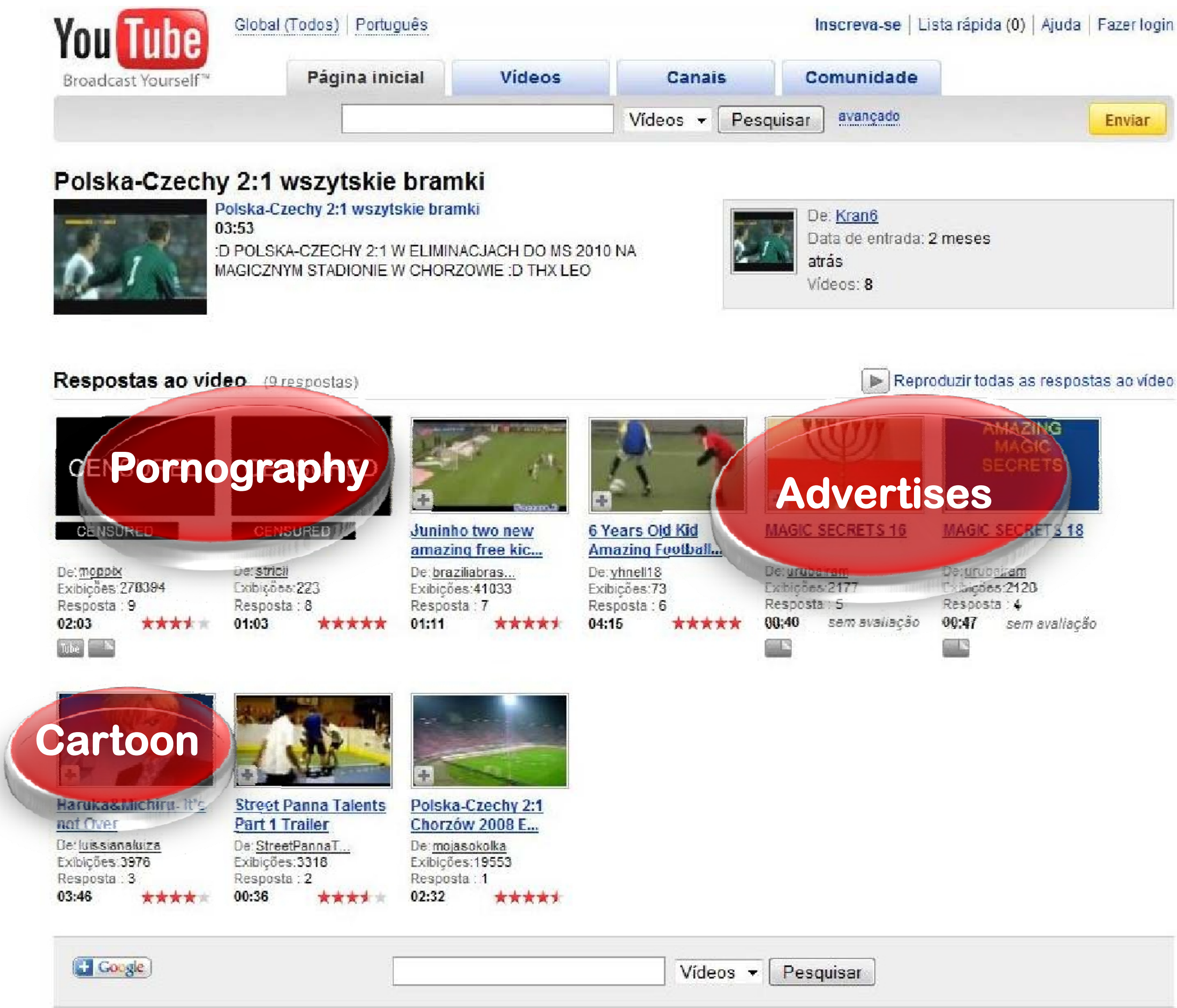
Detecting Video Spammers and Promoters

- Spammers
  - post an unrelated video as response to a popular video
- Promoters
  - Try to gain visibility to a specific video by posting a large number of (potentially unrelated) responses

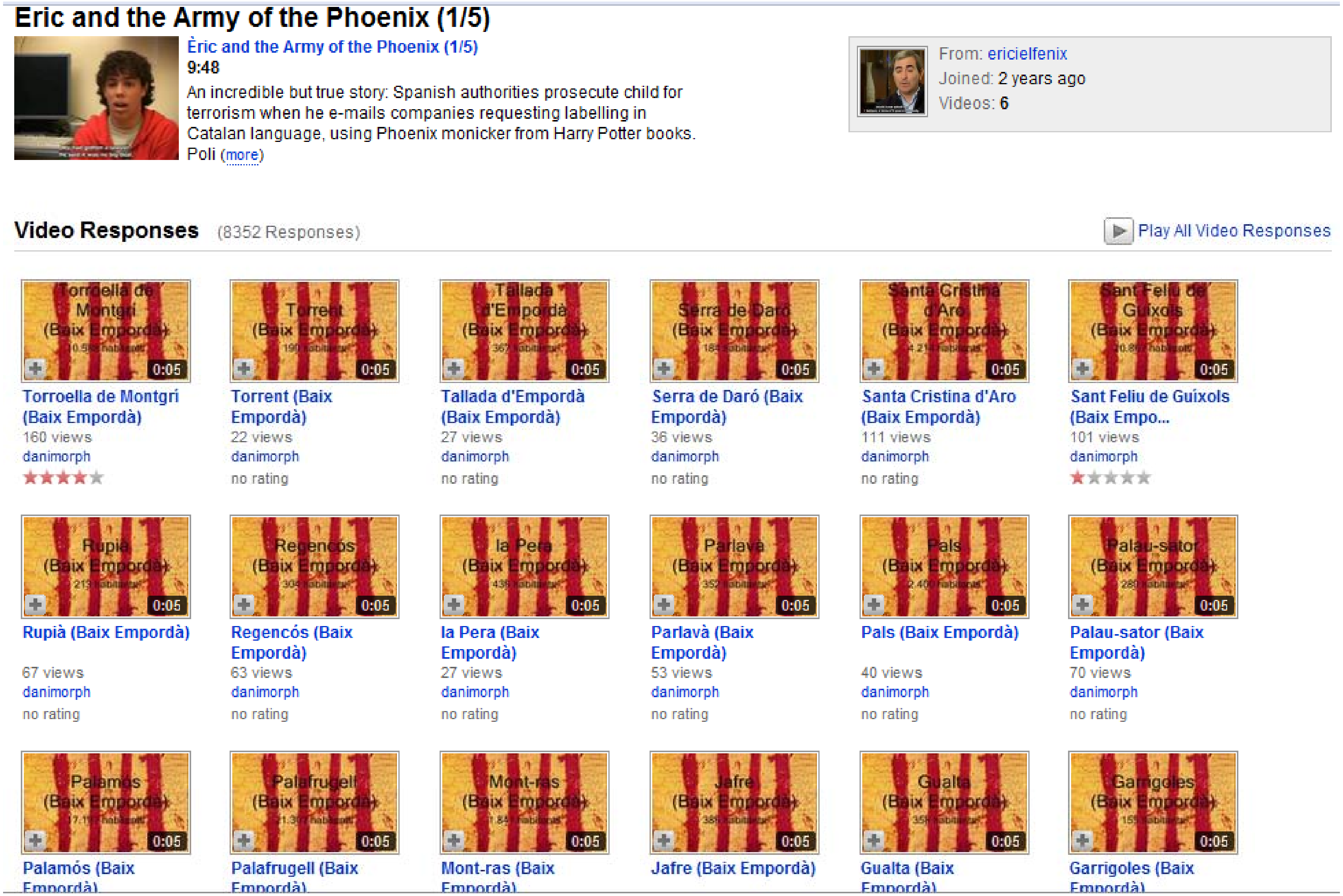
- 4-step approach
  1. Sample YouTube video responses and users
  2. Manually create a user test collection (promoters, spammers, and legitimate users)
  3. Identify attributes that can distinguish spammers and promoters from legitimate users
  4. Classification approach to detect spammers and promoters

Benevenuto, F., Rodrigues T., Almeida V., Almeida, J., and Gonçalves, M.
Detecting spammers and content promoters in online video social networks. In *SIGIR*, 2009.

Example of Video Spam



Example of Promotion



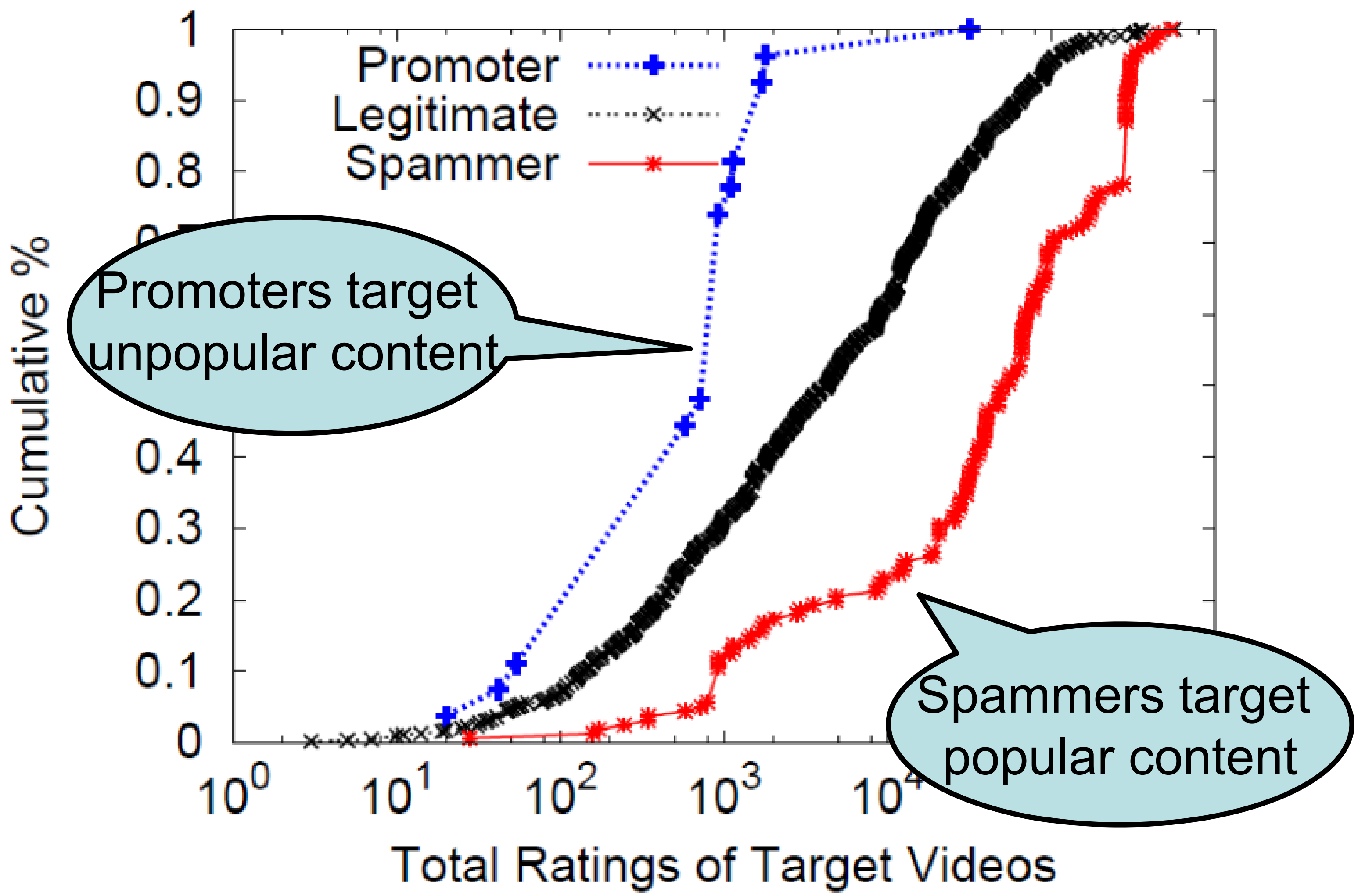
Step3. Attributes

- **User-Based:**
  - number of friends, number of subscriptions and subscribers, etc
- **Video-Based**:
  - duration, numbers of views and of comments received, ratings, etc
- **Social Network:**
  - clustering coefficient, betweenness, reciprocity, UserRank, etc

Feature Selection: χ^2 ranking

Attribute Set	Top 10	Top 20	Top 30	Top 40	Top 50
Video	9	18	25	30	36
User	1	2	4	7	9
SN	0	0	1	3	5

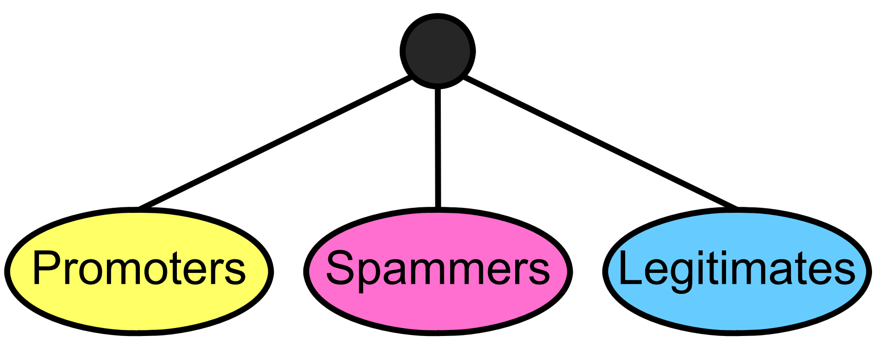
Distinguishing classes of users



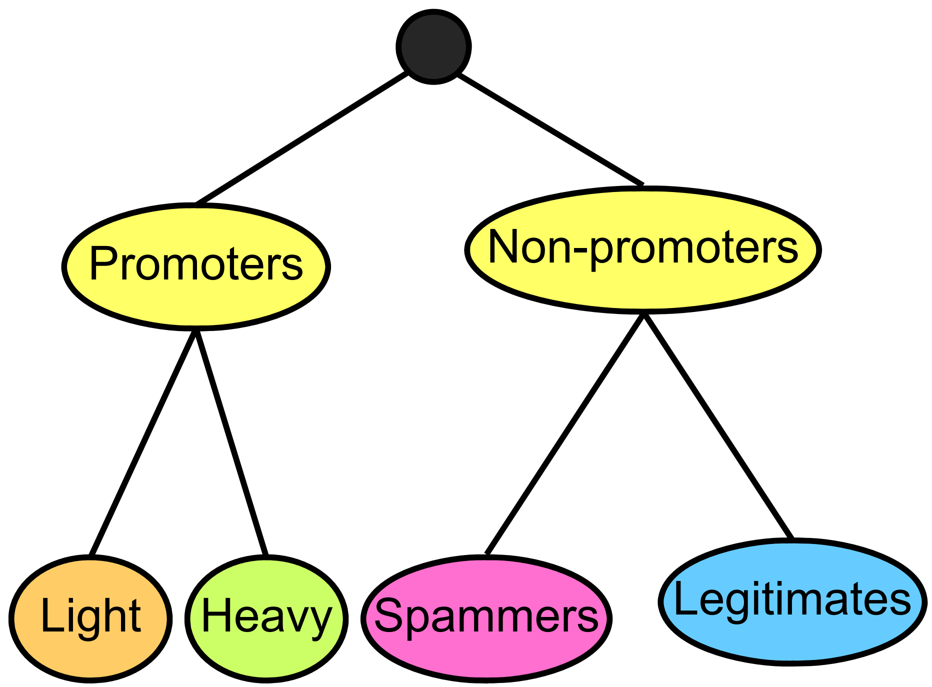
Step4. Classification Approach

- SVM (Support vector machine) as classifier
  - Use all attributes
  - Two classification approaches

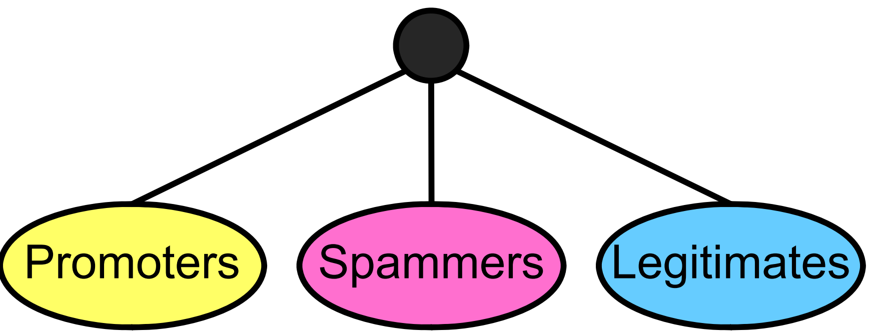
Flat



Hierarchical



Flat Classification

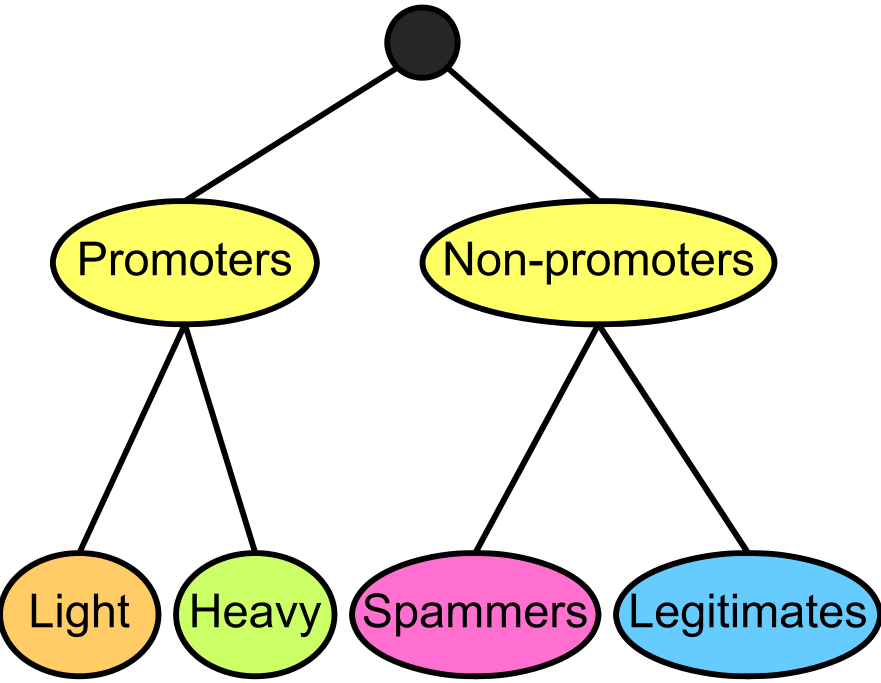


- Correctly identify majority of promoters, misclassifying a small fraction of legitimate users.
- Detect a significant fraction of spammers but they are much harder to distinguish from legitimate users.
  - Dual behavior of some spammers

		Predicted		
		Promoter	Spammer	Legitimate
True	Promoter	**96.13%**	3.87%	0.00%
	Spammer	1.40%	**56.69%**	41.91%
	Legitimate	0.31%	5.02%	**94.66%**

- Micro F1 = 88% (predict the correct class 88% of cases)

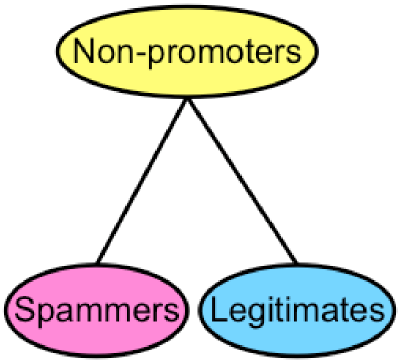
Hierarchical Classification



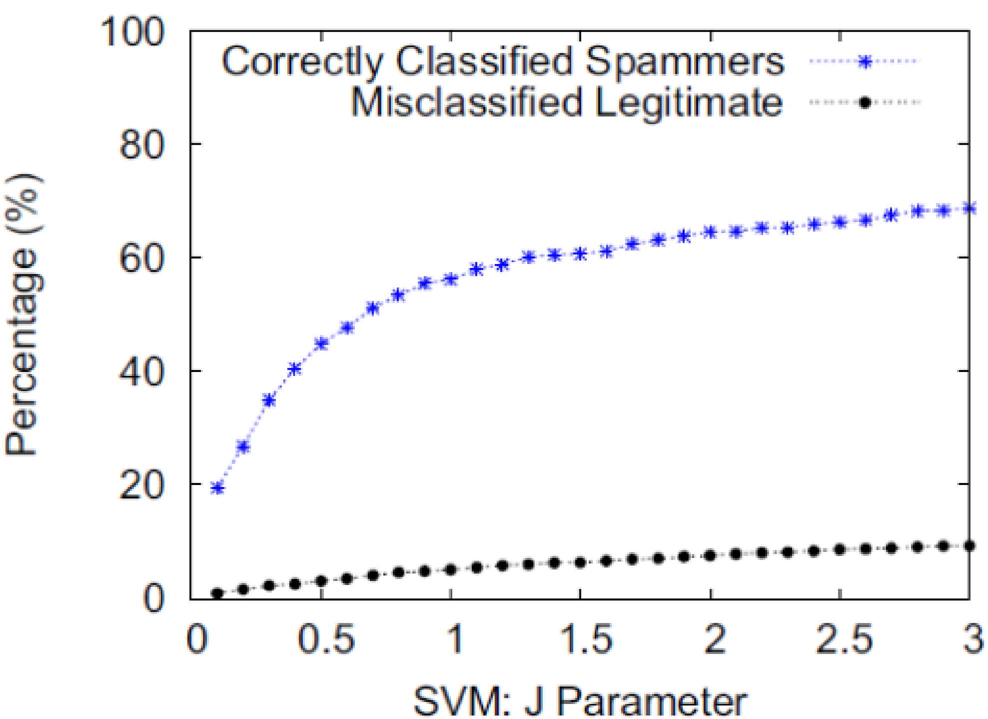
- Goal: provide flexibility in classification accuracy
- **First Level:**
  - Most promoters are correctly classified
  - Statistically indistinguishable compared with flat strategy

		Predicted	
		Promoter	Non-Promoter
	Promoter	**92.26%**	7.74%
True	Non-Promoter	0.55%	**99.45%**

Distinguishing Spammers from Legitimate users



		Predicted	
		Legitimate	Spammer
True	Legitimate	**95.09%**	4.91%
	Spammer	41.27%	**58.73%**



- **J = 0.1:** correctly classify 24% spammers, misclassifying <1% legitimate users
- **J = 3:** correctly classify 71% spammers, paying the cost of misclassifying 9% legitimate users

Foursqure Spam Tips

- Tips unrelated to Venue

Aggarwal, A., Almeida, J., and Kumaraguru, P. Detection of spam tipping behaviour on foursquare. In WWW Companion, 2013.

Features used to detect Spammers

- User Attributes
  - Properties of the Foursquare user profile and his checkins
- Social Attributes
  - Friends network of the Foursquare user under inspection
- Content Attributes
  - Details about Tips posted by the Foursquare user

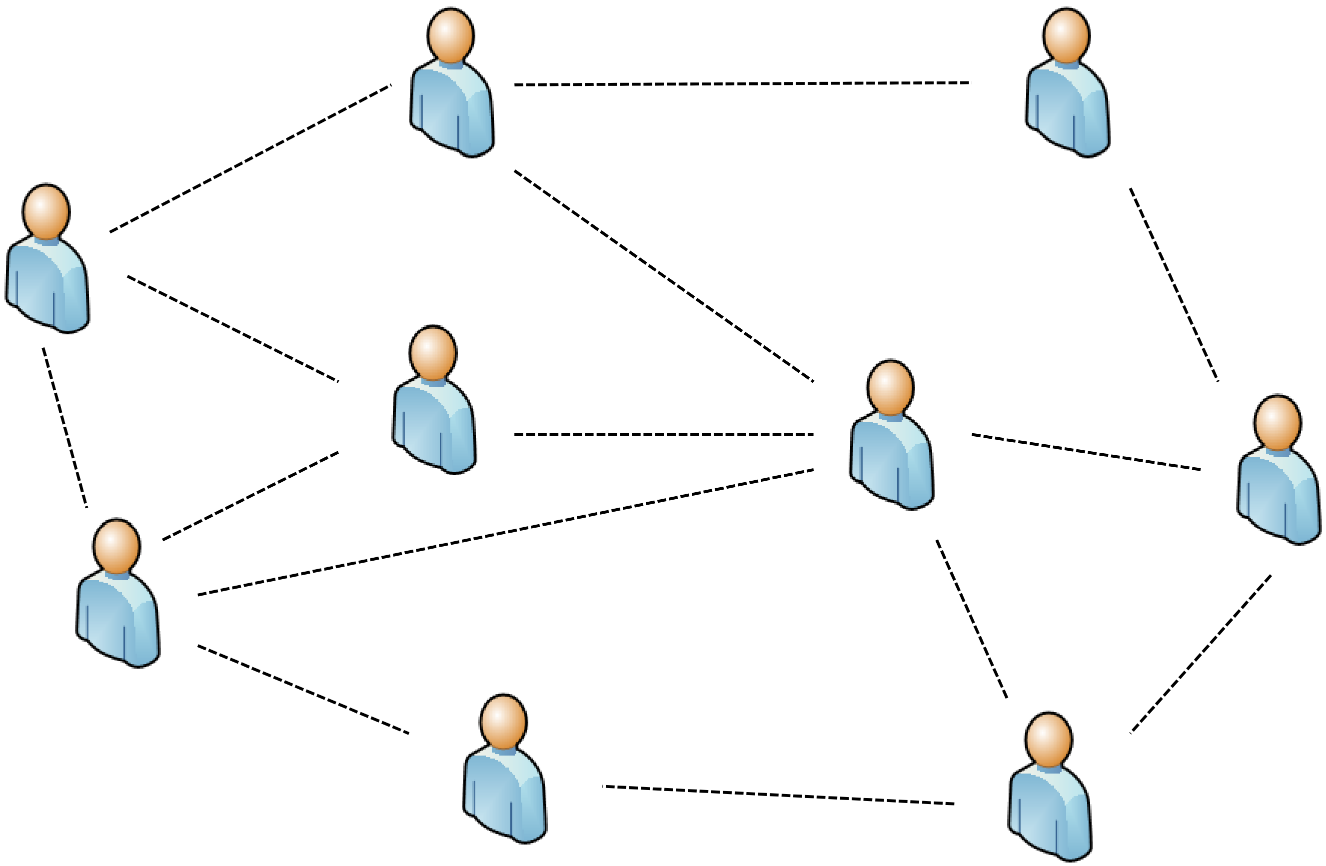
Features used

Category	χ2 rank	Feature
User Attributes	1	Number of Tips
	3	Ratio of Check-ins and Tips
	4	Number of Check-ins
	5	Number of Badges
	11	Number of Mayorships
	12	Ratio of Check-ins and Badges
	15	Number of Photos posted
Social Attributes	6	Number of Friends
Content Attributes	2	Similarity score of Tips
	7	Number of URLs posted
	8	Average number of words in Tips
	9	Average number of characters in Tips
	10	Ratio of number of likes and number of Tips
	13	Average number of spam words in Tips
	14	Average number of phone-numbers posted in Tips

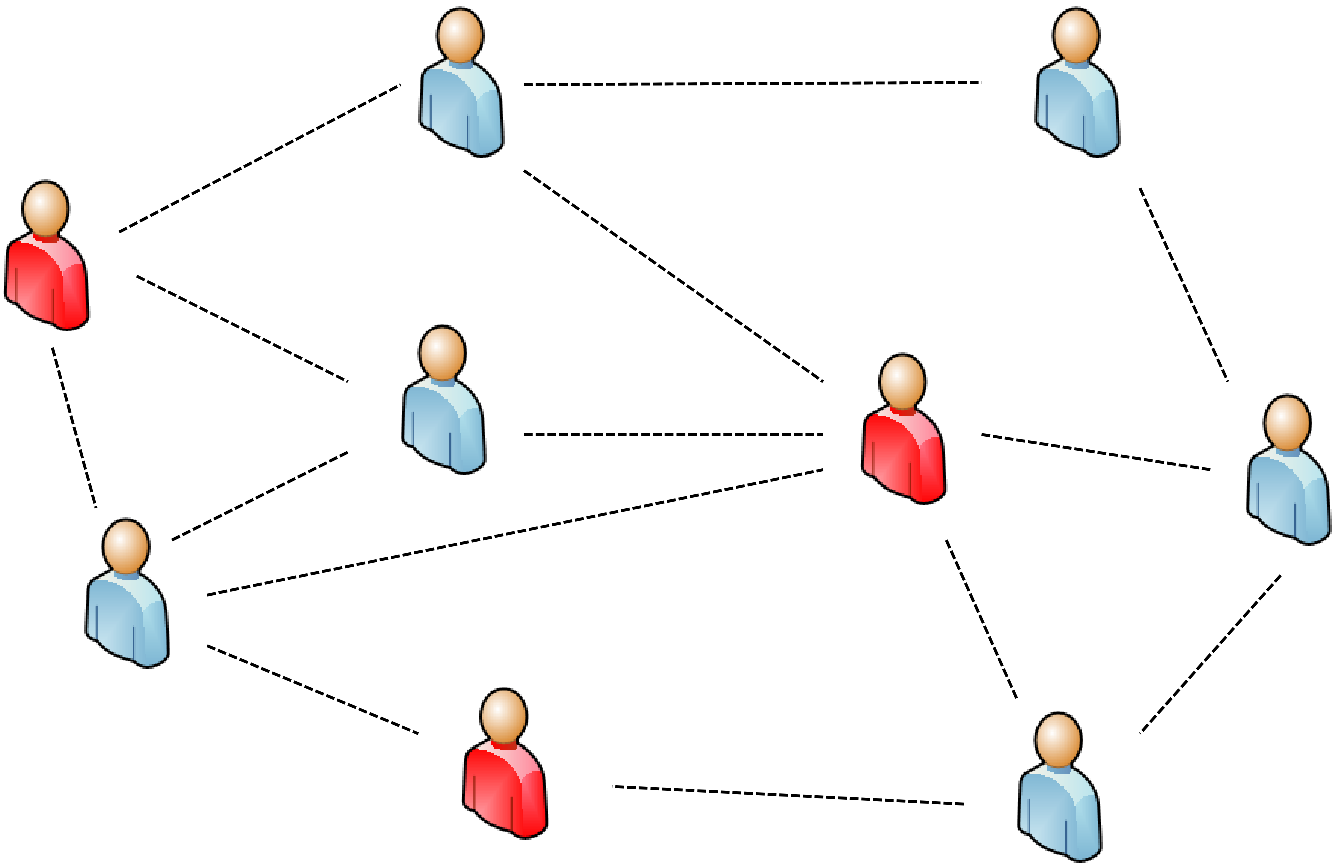
Classification Results

Classification Algorithm	Precision (Spam)	Precision (Safe)	Recall (Spam)	Recall (Safe)	Accuracy
KNN	83.2%	86.6%	86.3%	83.5%	84.89%
Decision Tree	88.1%	89.2%	88.3%	85.8%	89.53%
Random Forest	89.3%	90.2%	88.3%	90.3%	89.76%

How to Collect Evidence of Spammers

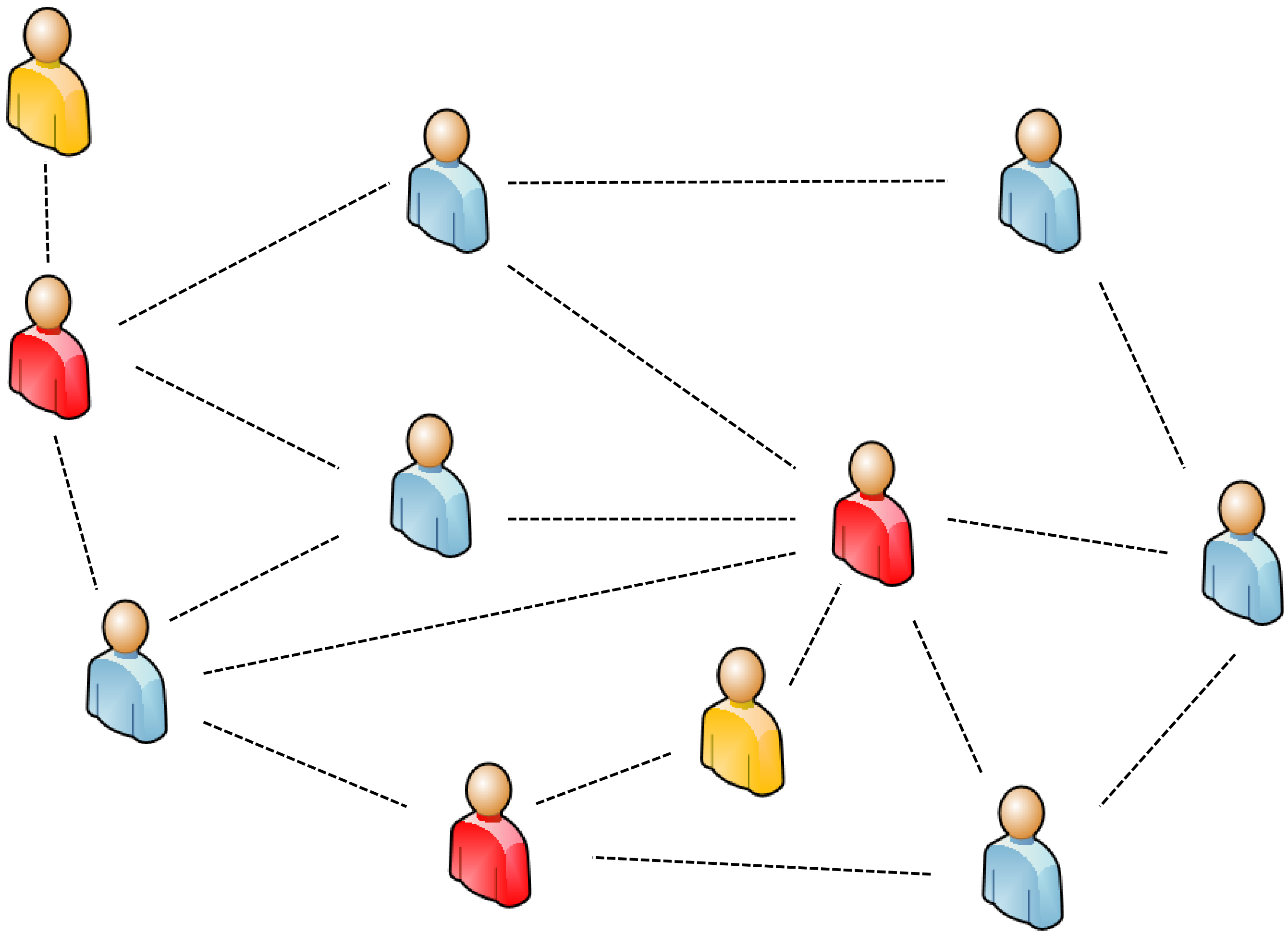


How to Collect Evidence of Spammers



- Human experts inspect users → Takes time to find spammers
- Users report spammers → 1) how many users participate? 2) False reports

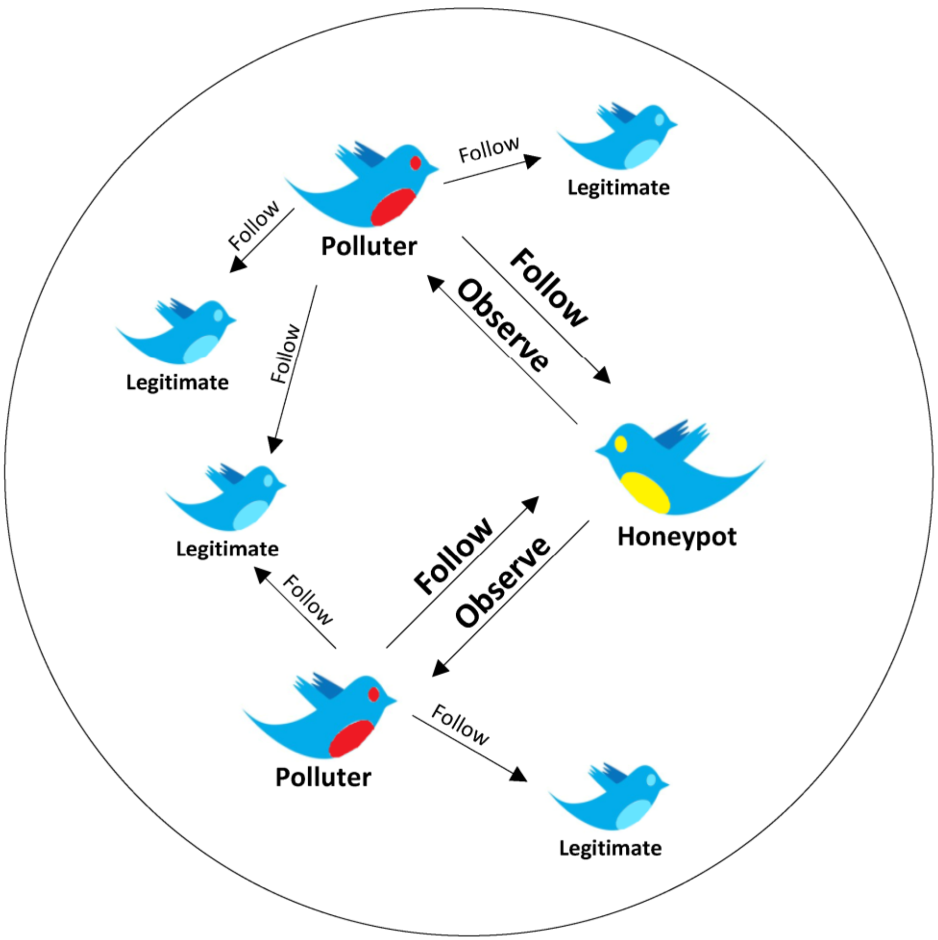
How to Collect Evidence of Spammers



- Create and deploy social honeypots in SNS

Lee, K., Eoff, B., and Caverlee, J. Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter. In *ICWSM*, 2011.

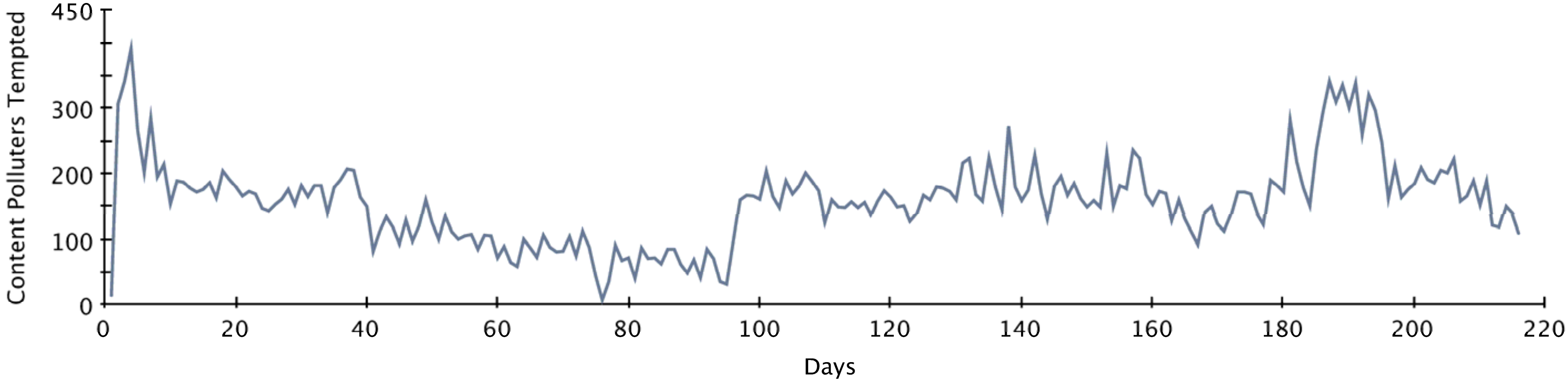
Social Honeypot Design



- Deployed 60 social honeypots (account + bot)
- They posted four types tweets with different ratio.
  - a normal textual tweet.
  - an "@" reply to one of the other social honeypots.
  - a tweet containing a link.
  - a tweet containing one of Twitter's current Top 10 trending topics, which are popular n-grams.
- Tempted 36,000 content polluters for seven months.

Study of Harvested Content Polluters

- The number of content polluters tempted per day

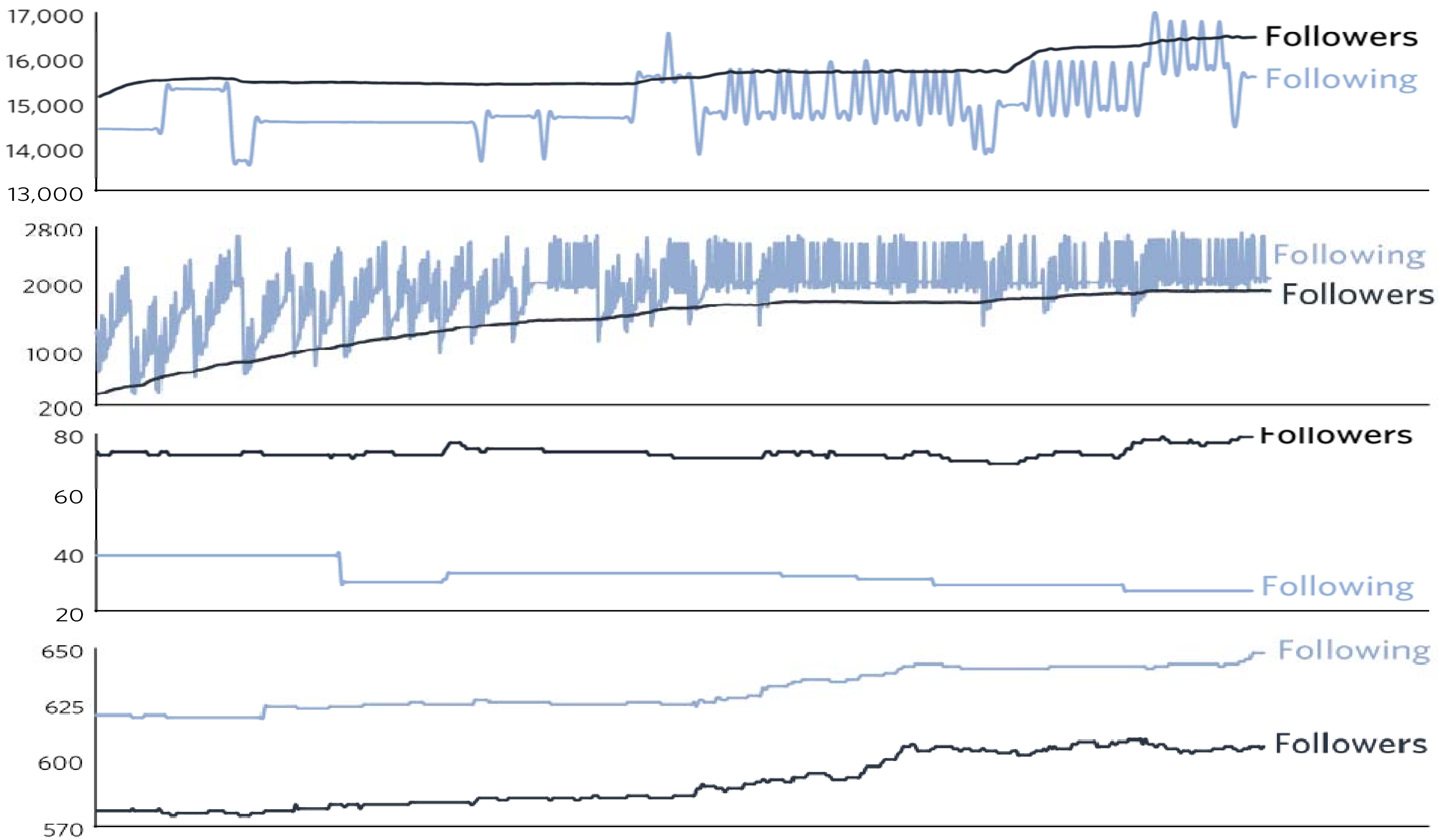


- Content Polluter Examples

Content Polluters	Examples
Duplicate Spammers	OFFICIAL PRESS RELEASE Limited To 10,000 "Platinum Founders" Reseller Licenses http://tinyurl.com/yd75xyy
Duplicate @ Spammers	#Follow @ anhran @PinkySparky @RestaurantsATL @combi31 @BBoomsma @TexMexAtl @DanielStoicaTax
Malicious Promoters	The Secret To Getting Lots Of Followers On Twitter http://bit.ly/6BiLk3
Friend Infiltrators	Thank you for the follows, from a newbie

Study of Harvested Content Polluters (Cont'd)

- Following and follower graphs of two content polluters and two legitimate users.



Ranking users based on their social graph

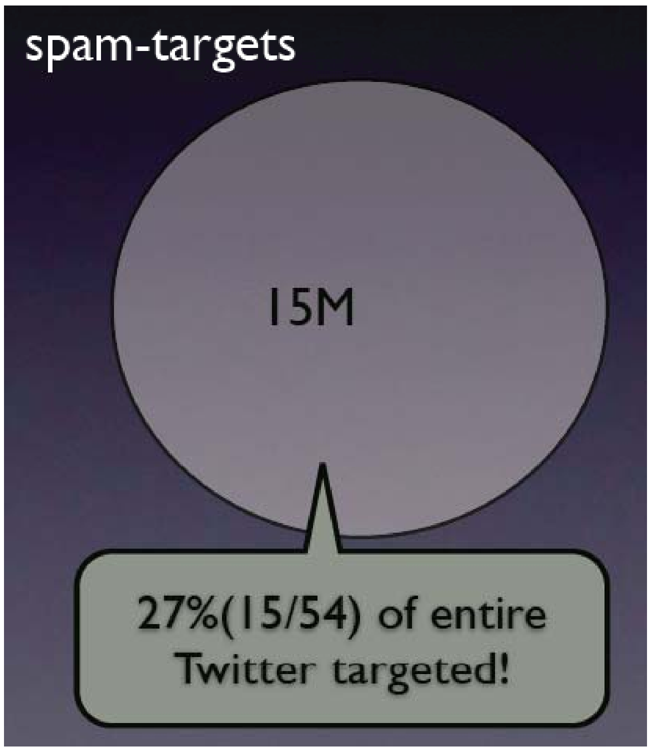
Identifying spammers

- Collected 54M Twitterusers, 1.9B links, 1.7B Tweets in 2009
- Identified the suspended accounts according to Twitter
  - Account could be suspended for various reasons
- Identified suspended users with at least one blacklisted URL
  - Includes 41,352 spammers

Ghosh, S., Viswanath, B., Kooti, F., Sharma, N. K., Korlam, G., Benevenuto, F., Ganguly, N., and Gummadi, P. K. Understanding and combating link farming in the twitter social network. In *WWW*, 2012.

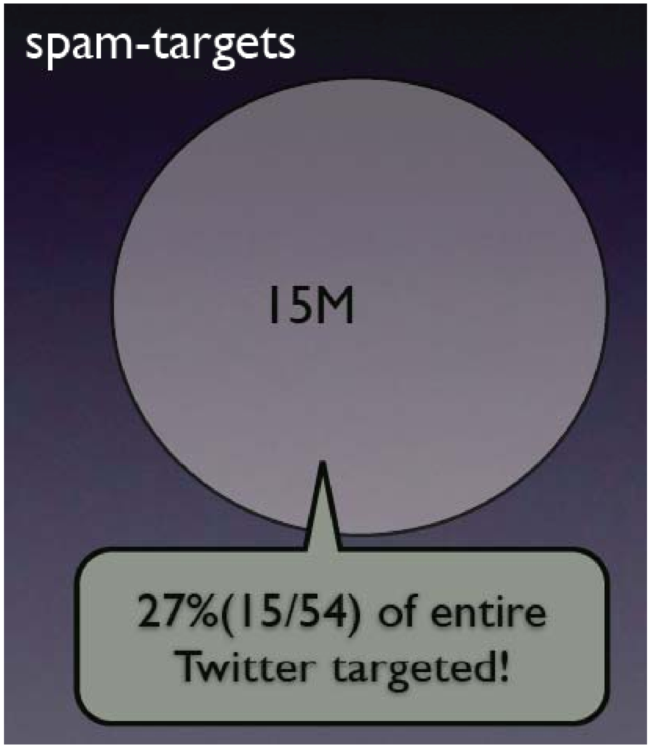
Do spammers engage in link farming?

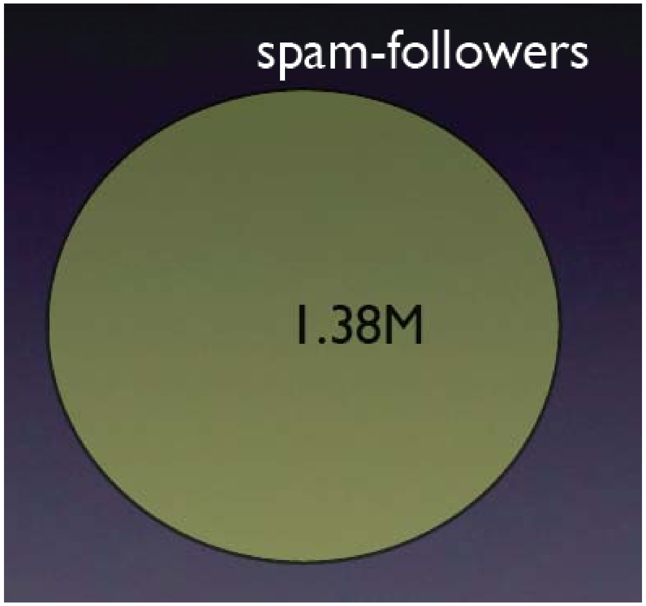
Spam-targets: Users followed by spammers



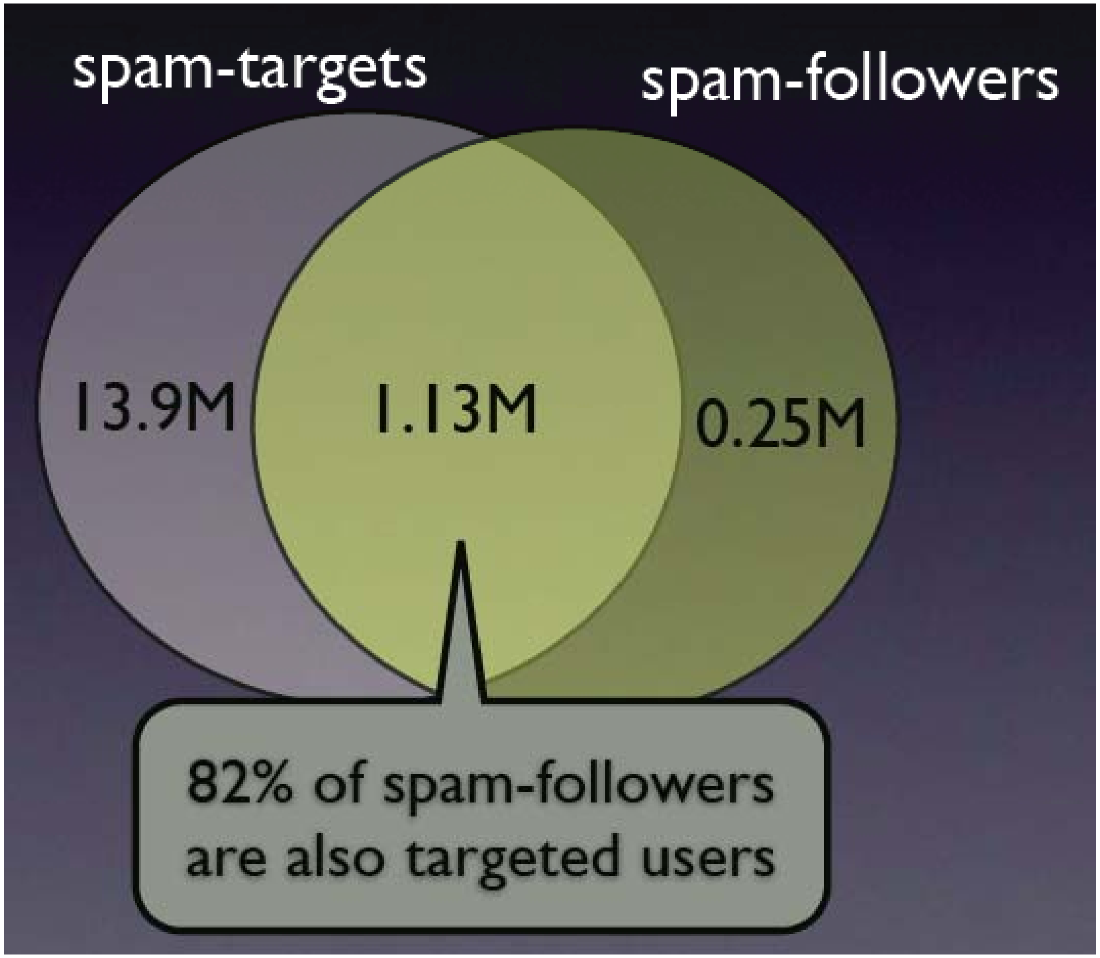
Do spammers engage in link farming?

Spam-followers: Users following spammers





Do spammers engage in link farming?



Spammers farm links at large-scale

Follower count for spammers is much higher than random users. Avg follower count for:
Spammers: 234,
Random users: 36

Are link farmers real users or spammers?

- To find out if they are spammers or real users, the reseracherss
  - 1. Used Twitter service to get list of suspended and verified users
    - 76% users not suspended, 235 of them verified by Twitter
  - 2. Manually verified 100 random users
    - 86% users are real with legitimate links in their Tweets
  - 3. Analyzed their profiles
    - They are much more active in updating their profiles than random users

- Link farmers are real active users

Who are the link farmers?

- Link farmers are mostly interested in promoting their business or tweeting about trends in a particular domain

Who are the link farmers?

- Top 5 link farmers according to Pagerank:
- 1. Barack Obama: Obama 2012 campaign staff
- 2. Britney Spears
- 3. NPR Politics: Political coverage and conversation
- 4. UK Prime Minister: PM’s office
- 5: JetBlue Airways

Link farmers include popular users and organizations

Collusionrank

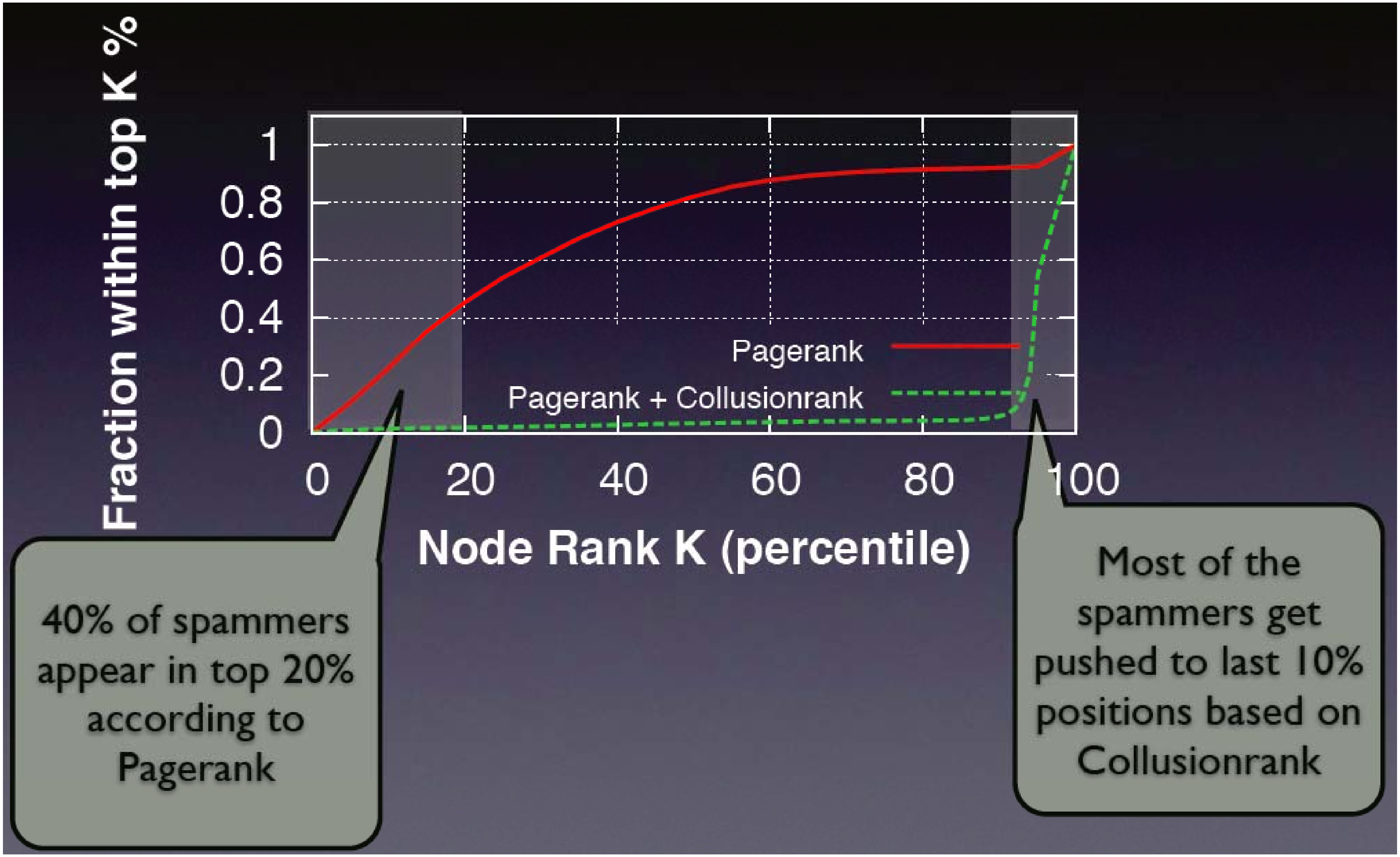
Algorithm:

- 1. Negatively bias the initial scores to the set of spammers
- 2. In Pagerank style, iteratively penalize users
  - who follow spammers or those who follow spam-followers

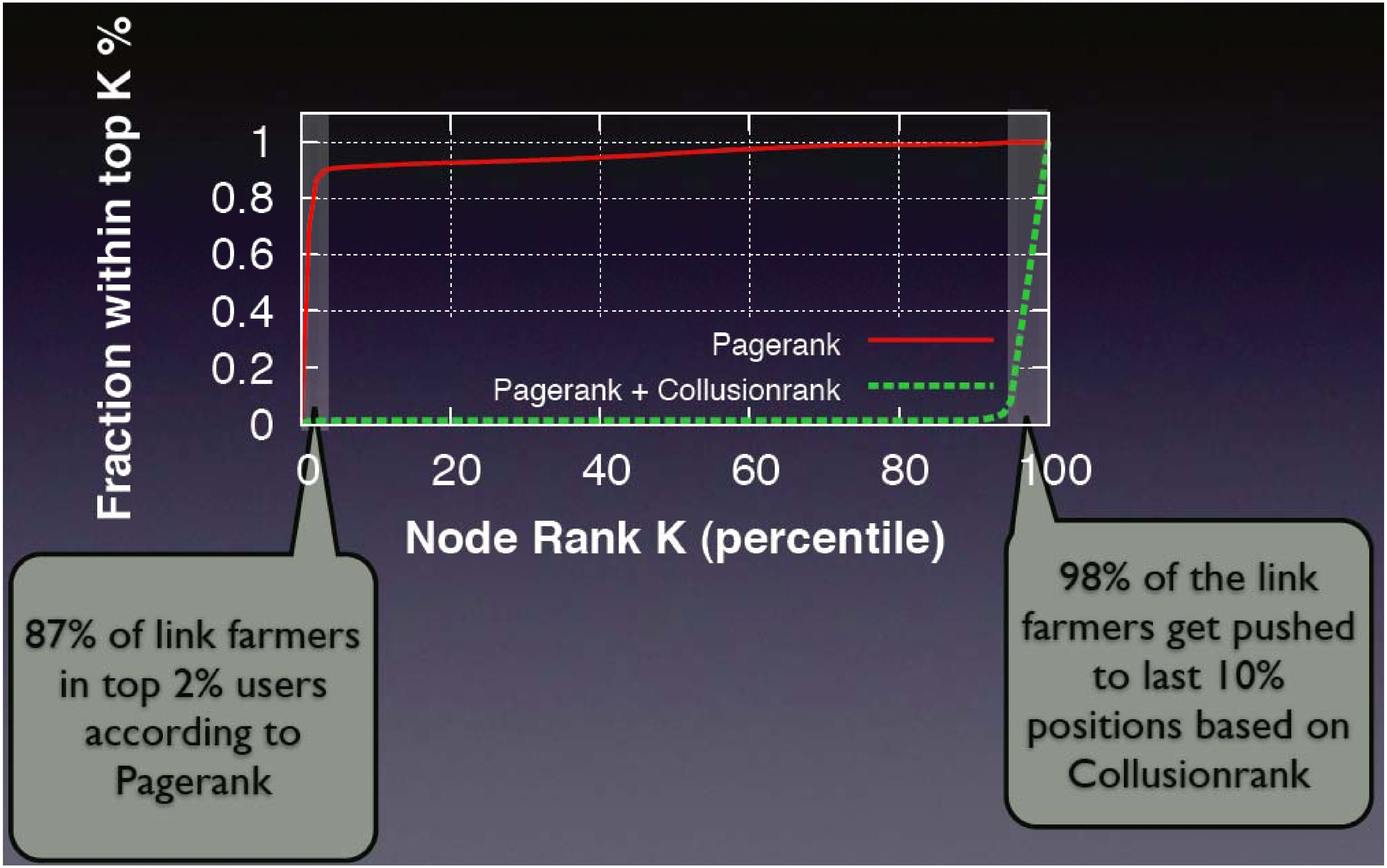
Collusionrank is based on the score of followings of a user

- Because user is penalized based on who he follows

Effect of Collusionrank on spammers



Effect on link farmers

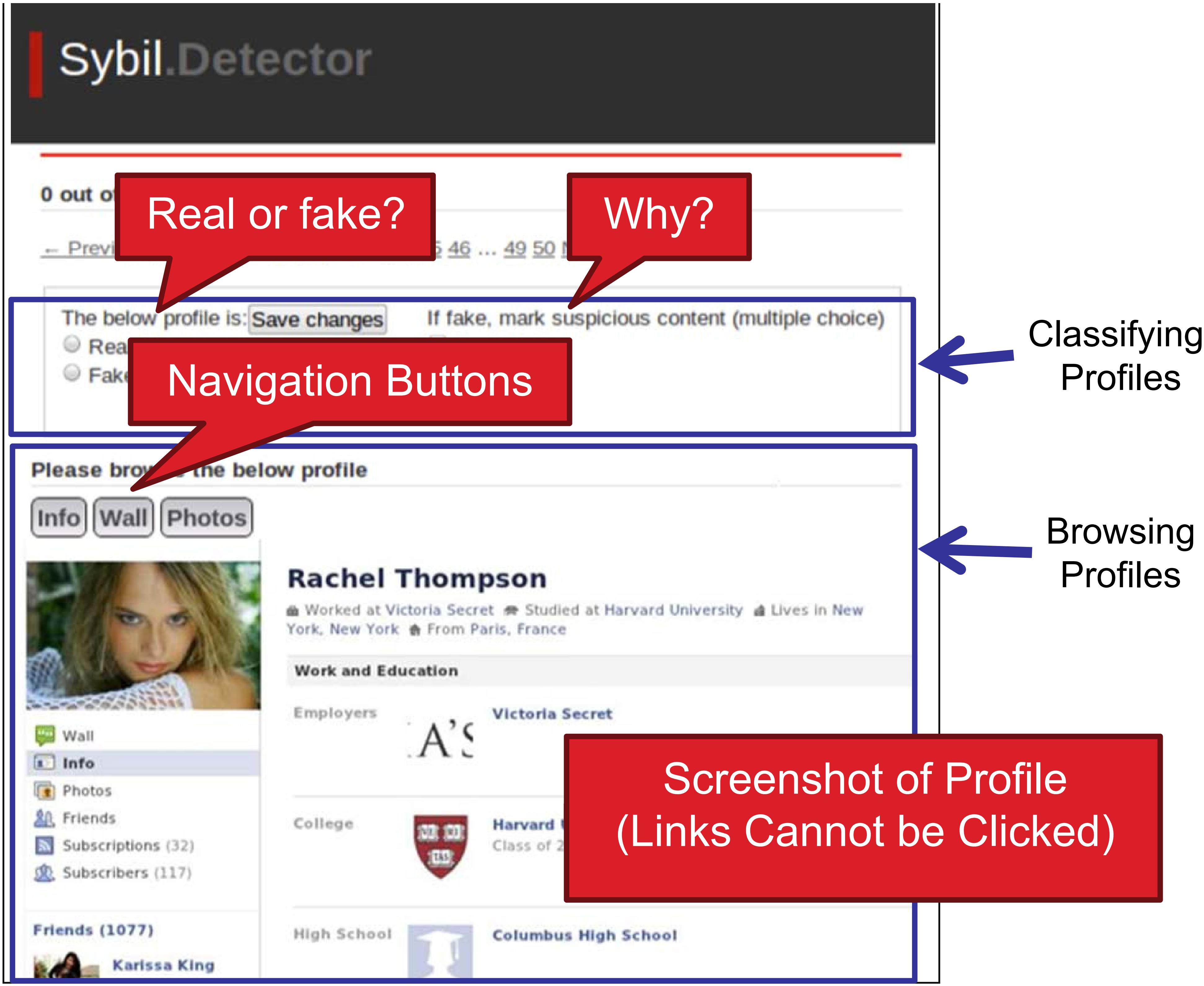


Using crowd wisdom (humans) to identify fake accounts (sybils)

User Study Setup

- User study with 2 groups of testers on 3 datasets
- 2 groups of users
  - Experts – The researchers' friends (CS professors and graduate students)
  - Turkers – Crowdworkers from online crowdsourcing systems
- 3 ground-truth datasets of full user profiles
  - Renren – given to them by Renren Inc.
  - Facebook US and India – crawled
    - Sybils (fake) profiles – banned profiles by Facebook
    - Legitimate profiles – 2-hops from the researchers' profiles

Wang, G., Mohanlal, M., Wilson, C., Wang, X., Metzger, M. J., Zheng, H., and Zhao, B. Y. Social Turing Tests: Crowdsourcing Sybil Detection. In NDSS, 2013.

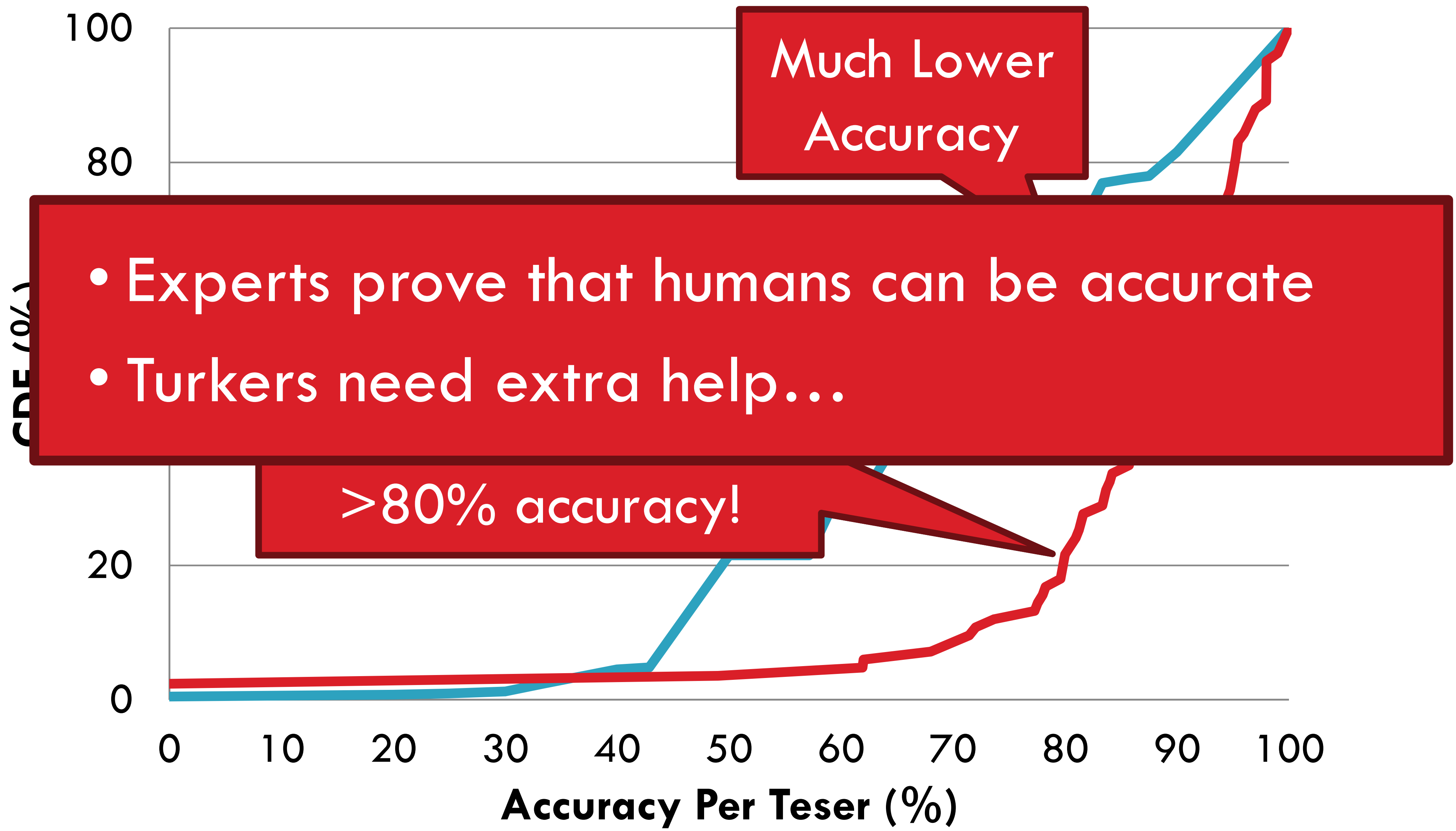


Sybil.Detector
0 out o
← Previ
46 … 49 50
Real or fake?
Why?
The below profile is: Save changes
If fake, mark suspicious content (multiple choice)
Rea
Fak
Navigation Buttons
Classifying Profiles
Please bro the below profile
Info
Wall
Photos
Browsing Profiles
Rachel Thompson
Worked at Victoria Secret
Studied at Harvard University
Lives in New York, New York
From Paris, France
Work and Education
Employers
Victoria Secret
College
Harvard
Class of 2
High School
Columbus High School
Wall
Info
Photos
Friends
Subscriptions (32)
Subscribers (117)
Friends (1077)
Karissa King
Screenshot of Profile (Links Cannot be Clicked)

Experiment Overview

Dataset	# of Profiles		Test Group	# of Testers	Profile per Tester
	Sybil	Legit.			
Renren	100	100	Chinese Expert	24	100
			Chinese Turker	418	10
Facebook US	32	50	US Expert	40	50
			US Turker	299	12
Facebook India	50	49	India Expert	20	100
			India Turker	342	12

Individual Tester Accuracy



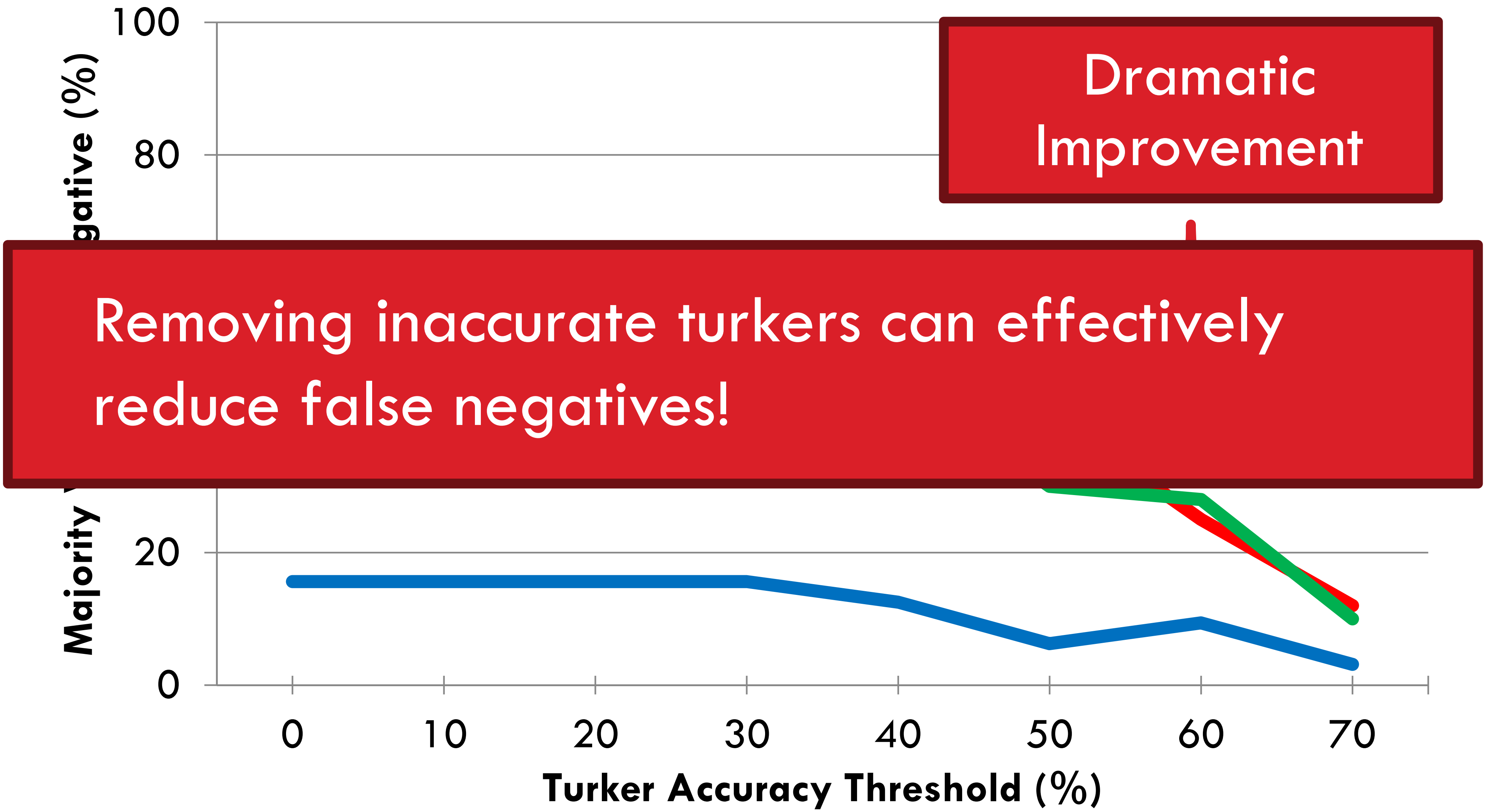
- Experts prove that humans can be accurate
- Turkers need extra help…

Wisdom of the Crowd

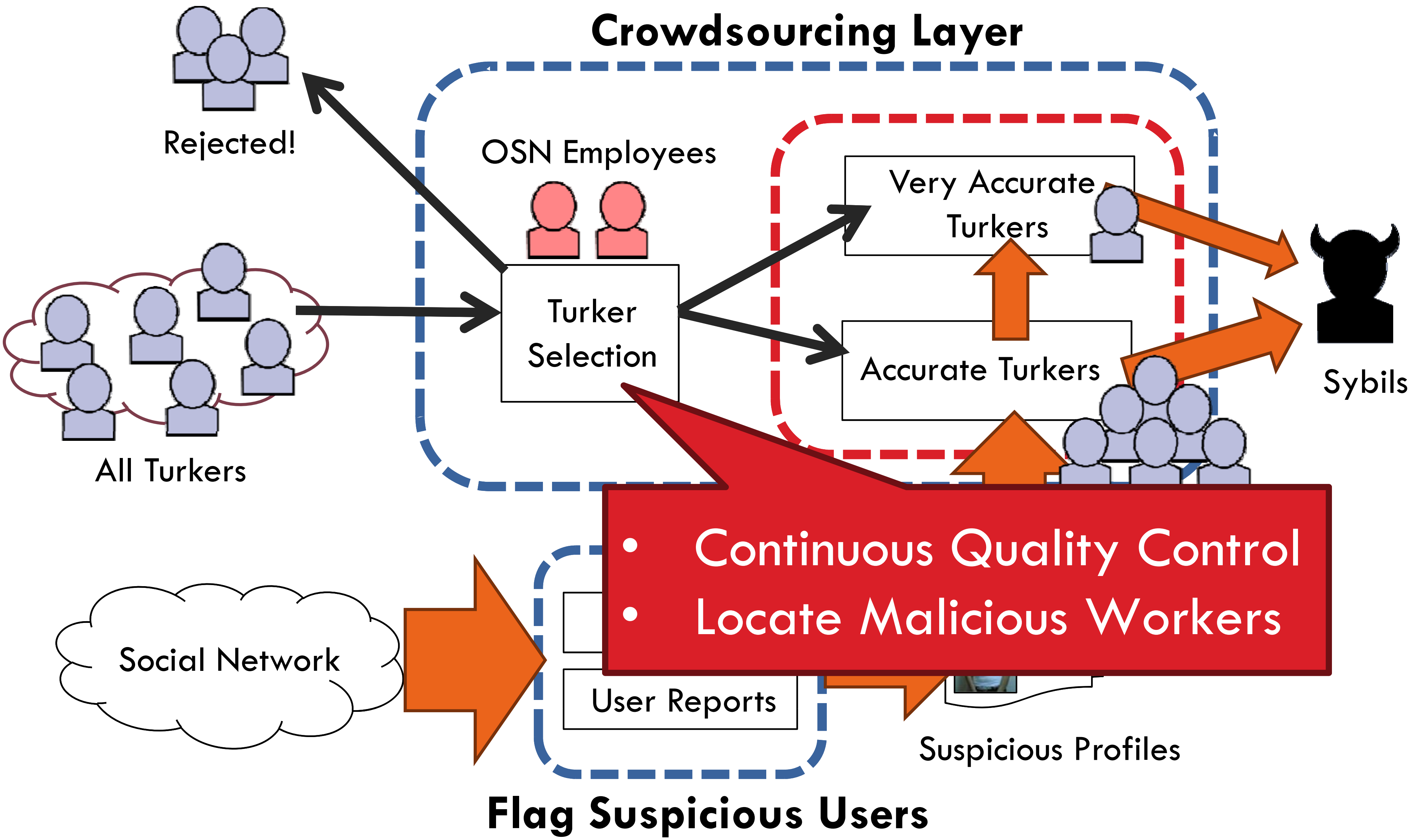
- Is wisdom of the crowd enough?
- Majority voting
  - Treat each classification by each tester as a vote
  - Majority vote determines final decision of the crowd
- Results after majority voting (20 votes)

- False positive rates are excellent
- What can be done to improve turker accuracy?

Eliminating Inaccurate Turkers



System Architecture



So far… Social Spam Detection Approaches

- Supervised spam detection approach
  - The most popular approach
  - Require labeled data for training purpose

- Ranking users based on their social graph

- Use crowd wisdom (humans) to identify fake accounts

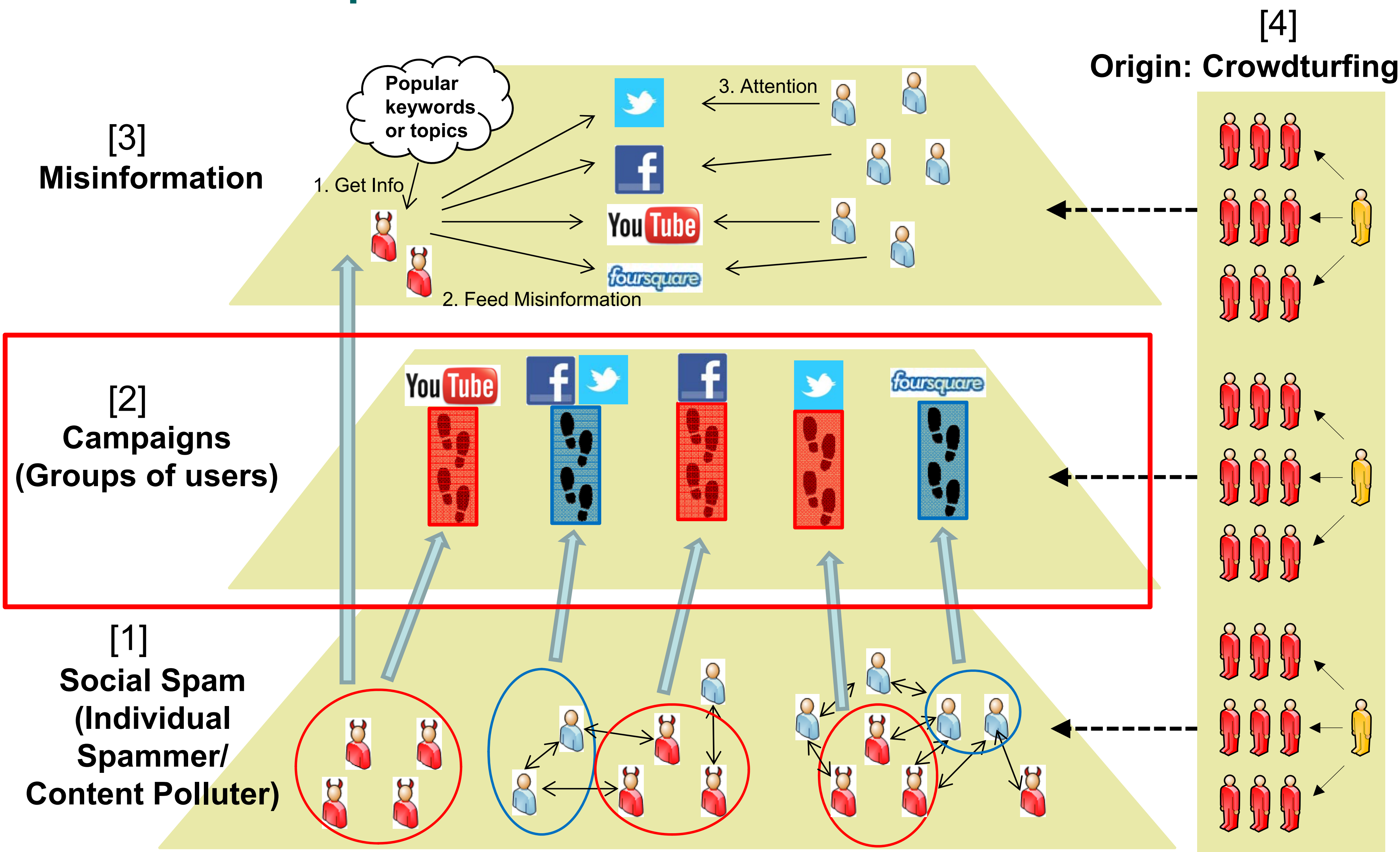
Reference List

- Grier, C., Thomas, K., Paxson, V., and Zhang, M. @spam: the underground on 140 characters or less. In CCS, 2010.
- Lee, S., and Kim, J. WarningBird: Detecting suspicious URLs in Twitter stream. In NDSS, 2012.
- Ghosh, S., Viswanath, B., Kooti, F., Sharma, N. K., Korlam, G., Benevenuto, F., Ganguly, N., and Gummadi, P. K. Understanding and combating link farming in the twitter social network. In WWW, 2012.
- Benevenuto, F., Rodrigues T., Almeida V., Almeida, J., and Gonçalves, M. Detecting spammers and content promoters in online video social networks. In SIGIR, 2009.
- Lee, K., Eoff, B., and Caverlee, J. Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter. In ICWSM, 2011.
- Aggarwal, A., Almeida, J., and Kumaraguru, P. Detection of spam tipping behaviour on foursquare. In WWW Companion, 2013.
- Lee., K., Caverlee., J., and Webb, S. Uncovering Social Spammers: Social Honeypots + Machine Learning. In SIGIR, 2010.
- Wang, G., Mohanlal, M., Wilson, C., Wang, X., Metzger, M. J., Zheng, H., and Zhao, B. Y. Social Turing Tests: Crowdsourcing Sybil Detection. In NDSS, 2013.
- Tan, E., Guo, L., Chen, S., Zhang, X., and Zhao, Y. UNIK: Unsupervised Social Network Spam Detection. In CIKM, 2013
- Lee, K., Kamath, K., and Caverlee, J. Combating Threats to Collective Attention in Social Media: An Evaluation. In ICWSM, 2013.

Schedule

14:00 ~ 14:10	Introduction to Social Media Threats (Social Spam, Campaigns, Misinformation and Crowdturfing)
14:10 ~ 14:55	Social Spam
14:55 ~ 15:30	Campaigns
15:30 ~ 16:00	Break
16:00 ~ 16:30	Misinformation
16:30 ~ 17:10	Crowdturfing
17:10 ~ 17:30	Challenges, Tools and Conclusion

Conceptual Level of Tutorial Theme



Campaign Detection Approaches

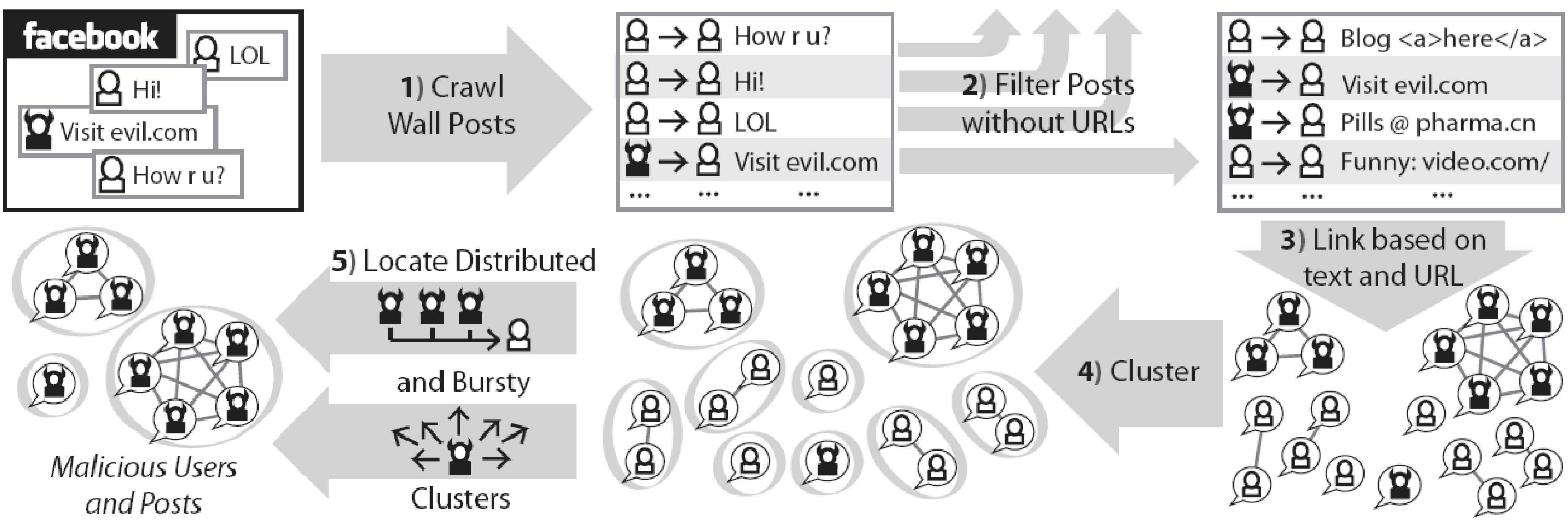
- Graph-based spam campaign detection

- Content-driven campaign detection

Graph-based spam campaign detection

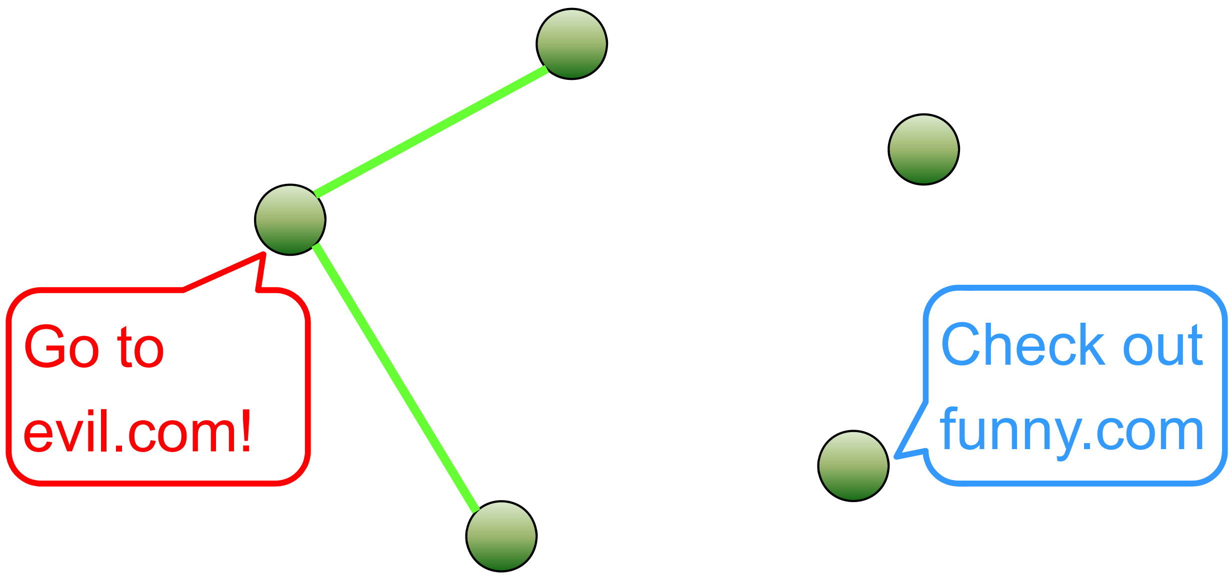
System Overview

- Identify coordinated spam campaigns in Facebook.
  - Templates are used for spam generation.



Gao, H., Hu J., Wilson, C., Li, Z., Chen, Y., and Zhao, B. Detecting and characterizing social spam campaigns. In IMC, 2010.

Build Post Similarity Graph



– A node: an individual wall post

– An edge: connect two “similar” wall posts

Wall Post Similarity Metric

Spam wall post model:

A textual description:

hey see your love compatibility ! go here yourlovecalc . com (remove spaces)

A destination URL:

Wall Post Similarity Metric

- Condition 1:
  - Similar textual description.

Guess who your secret admirer is??
Go here nevasubevd . blogs pot . co m **(take out spaces)**

Guess who your secret admirer is??”
Visit: yes crush com (remove spaces)

Establish an edge!

Wall Post Similarity Metric

- Condition 2:
  - Same destination URL.

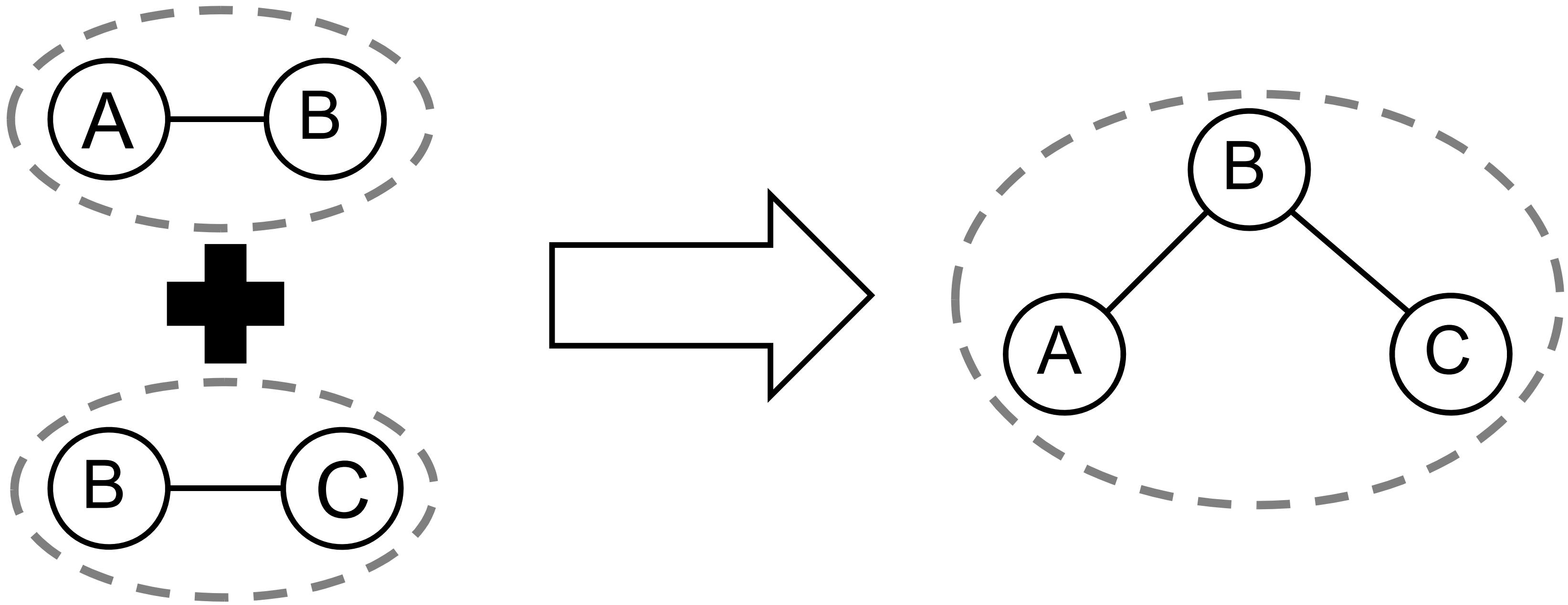
secret admirer revealed.
goto yourlovecalc . com (remove the spaces)

hey see your love compatibility !
go here yourlovecalc . com (remove spaces)

Establish an edge!

Extract Wall Post Campaigns

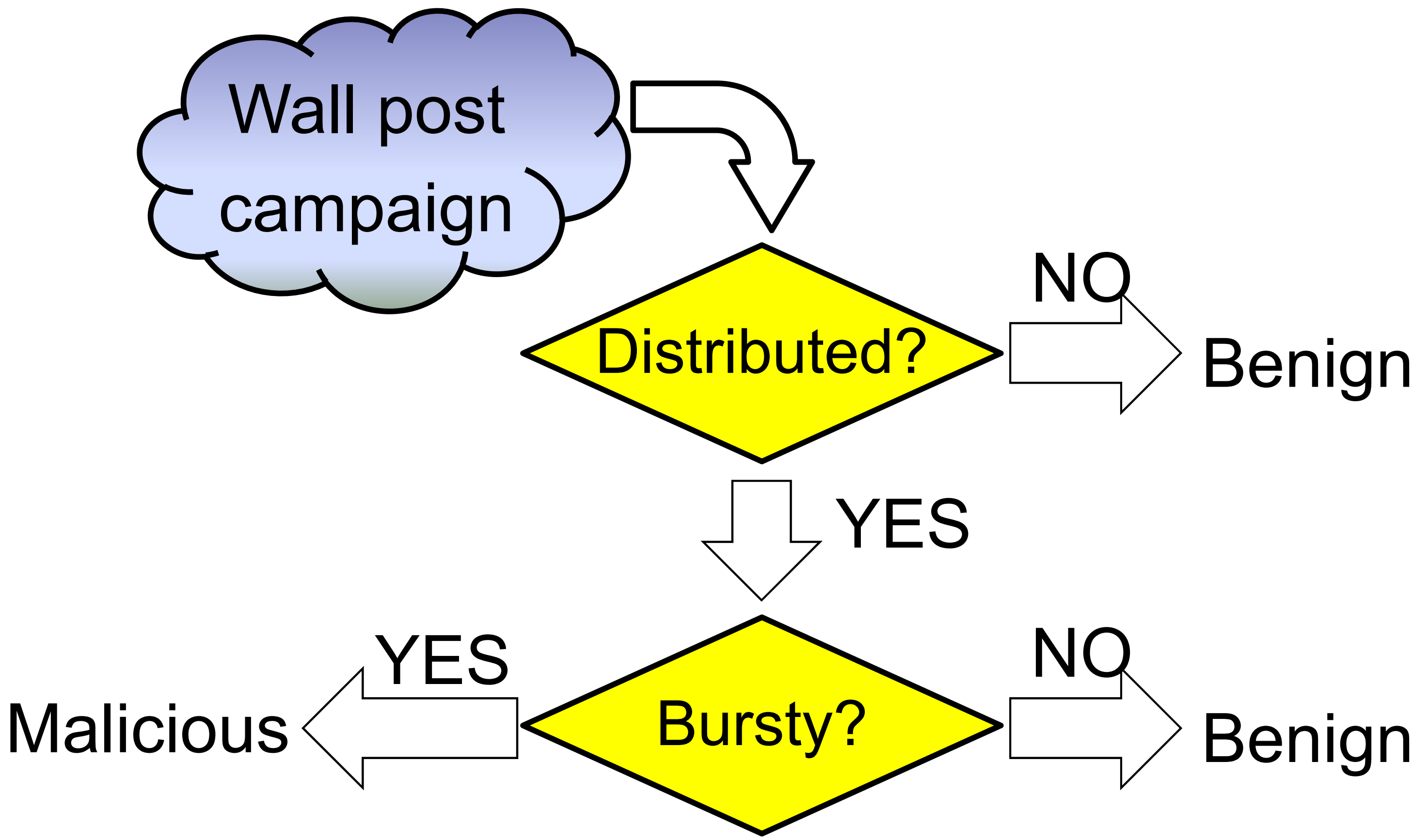
- Intuition:



- Reduce the problem of identifying potential campaigns to identifying connected subgraphs.

Locate Spam Campaigns

- Distributed: campaigns have many senders.
- Bursty: campaigns send fast.



Validation

- The detection approach found ~200K malicious wall posts (~10%) from ~2M wall posts with URLs.
- Validation focused on detected URLs.
- Adopted multiple validation steps:
  - URL de-obfuscation
  - Keyword matching
  - 3rd party tools
  - URL grouping
  - Redirection analysis
  - Manual confirmation

Validation

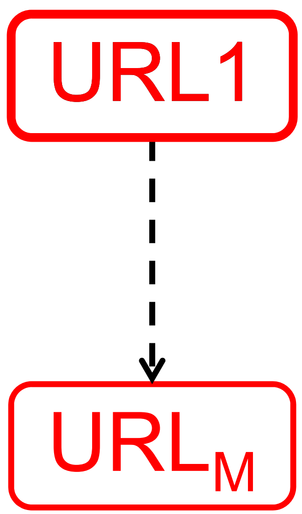
- Step 1: Obfuscated URL
  - URLs embedded with obfuscation are malicious.
  - Reverse engineer URL obfuscation methods:
    - Replace ‘.’ with “dot” : 1lovecrush dot com
    - Insert white spaces : abbykywyty . blogs pot . co m

Validation

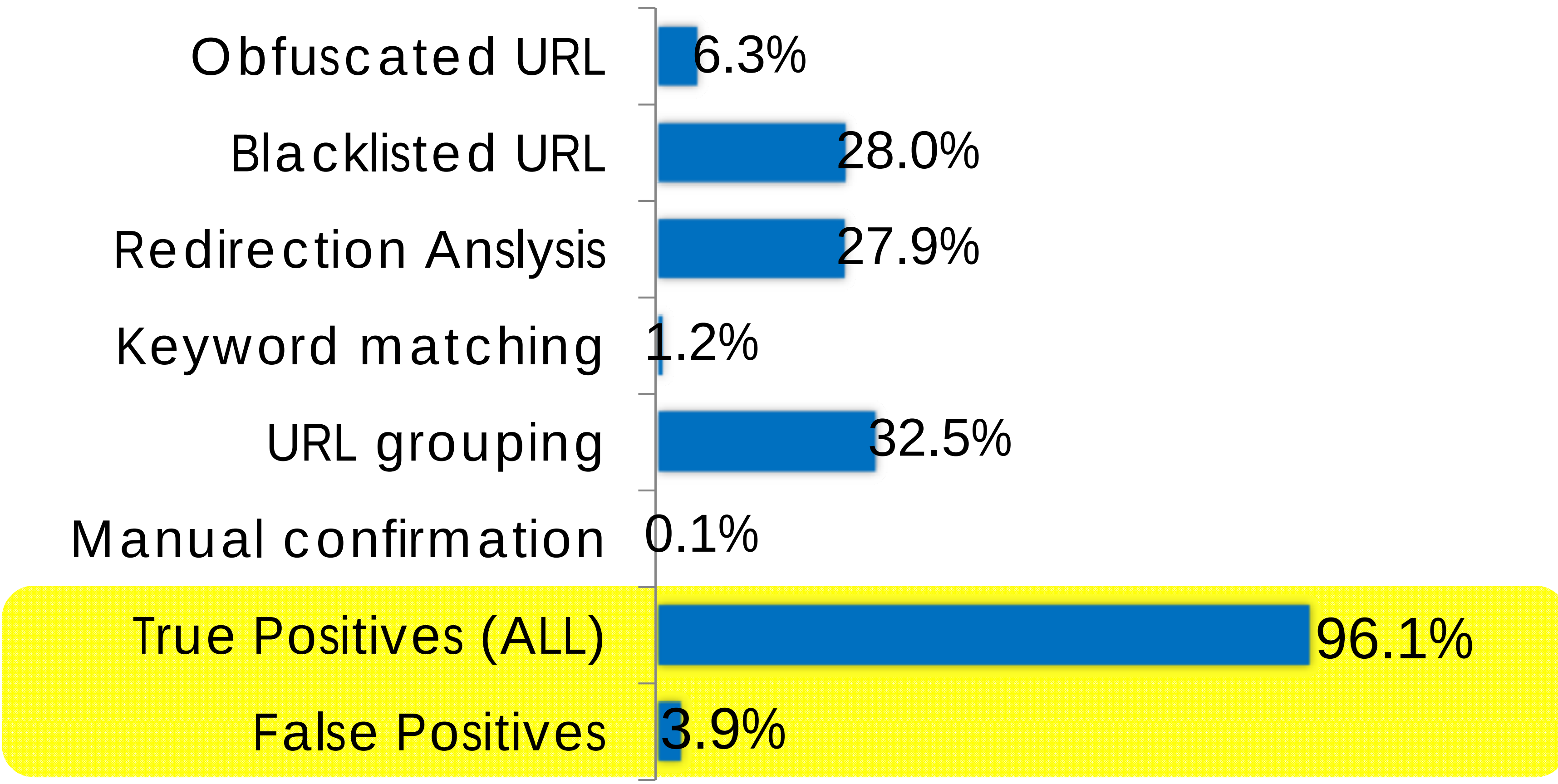
- Step 2: Third-party tools
  - Use multiple tools, including:
    - McAfee SiteAdvisor
    - Google's Safe Browsing API
    - Spamhaus
    - Wepawet (a drive-by-download analysis tool)
    - …

Validation

- Step 3: Redirection analysis
  - Commonly used by the attackers to hide the malicious URLs.

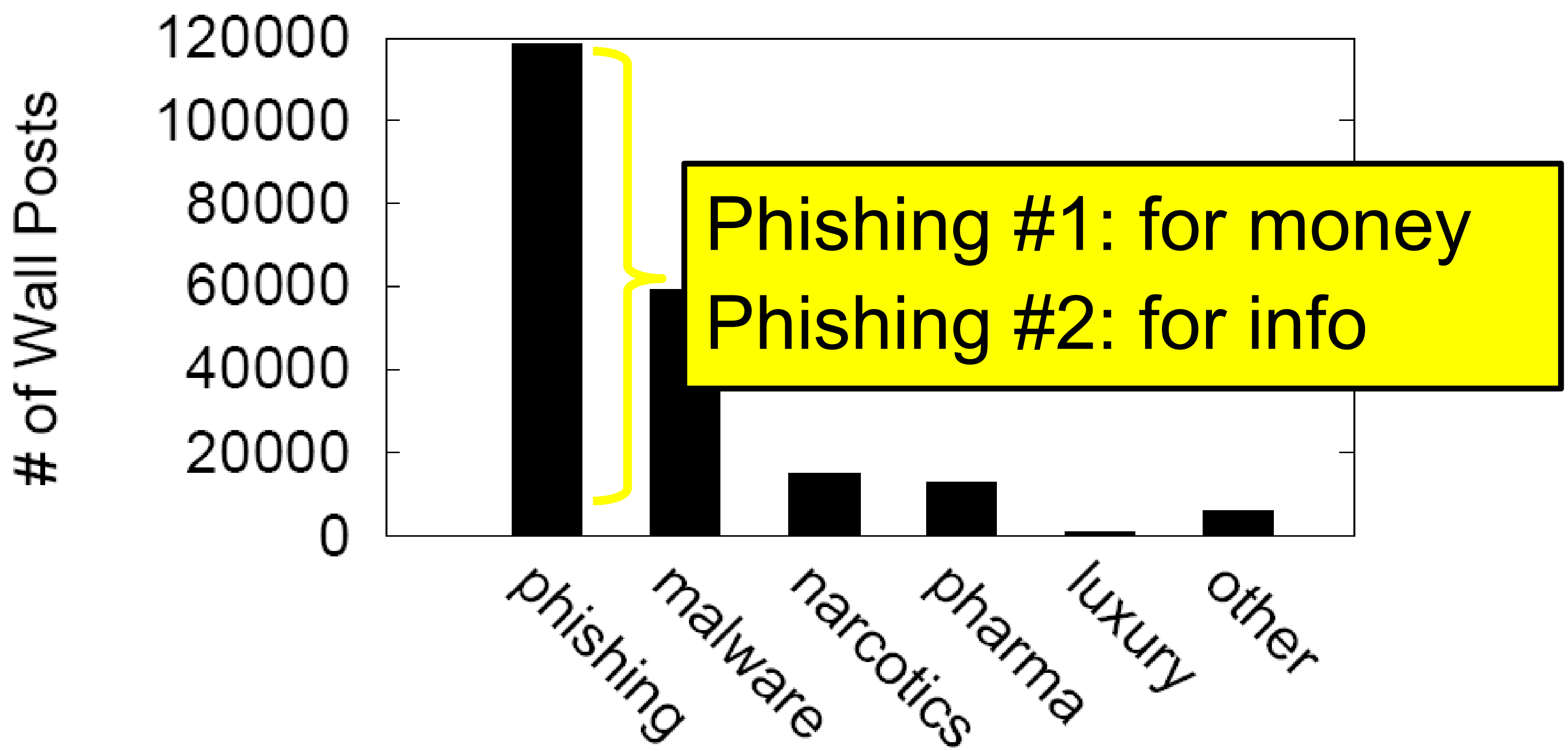


Experimental Evaluation



The validation result.

Spam Campaign Goal Analysis

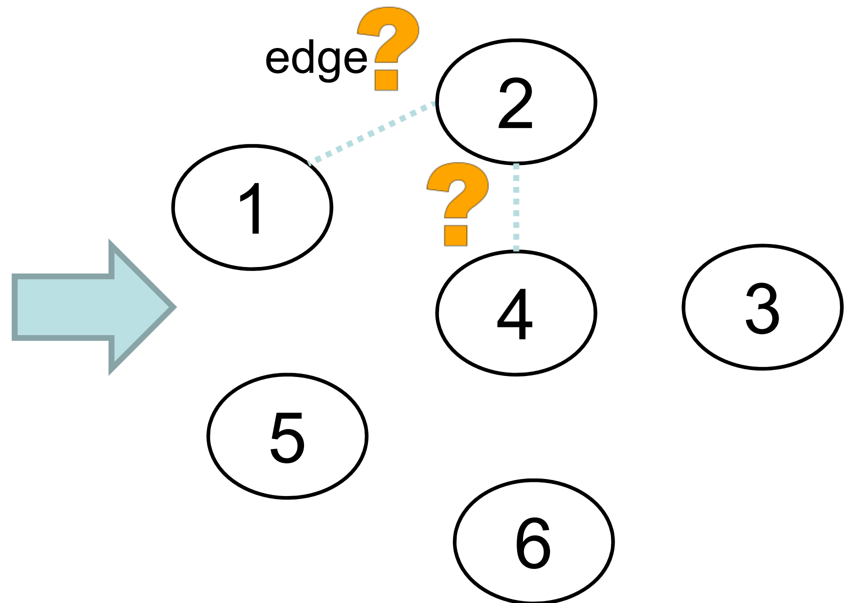


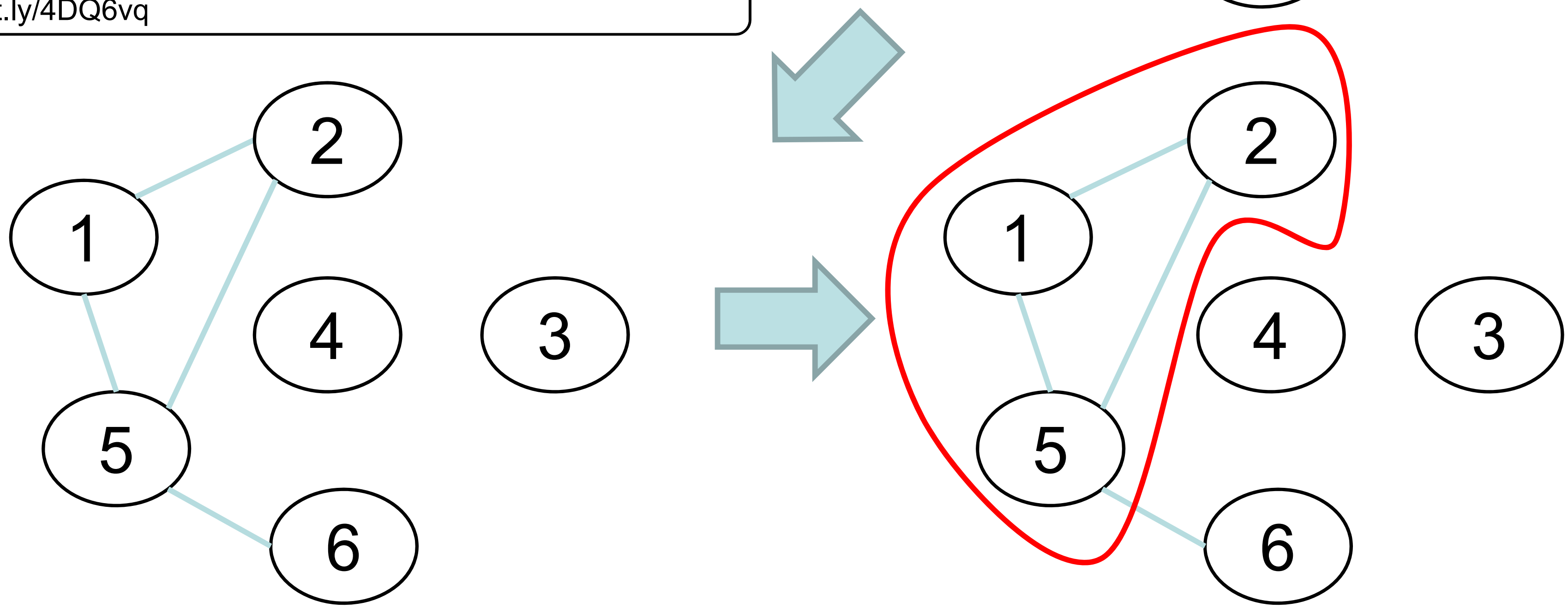
- Categorize the attacks by attackers' goals.

Content-driven campaign detection

Message Level Campaign Detection

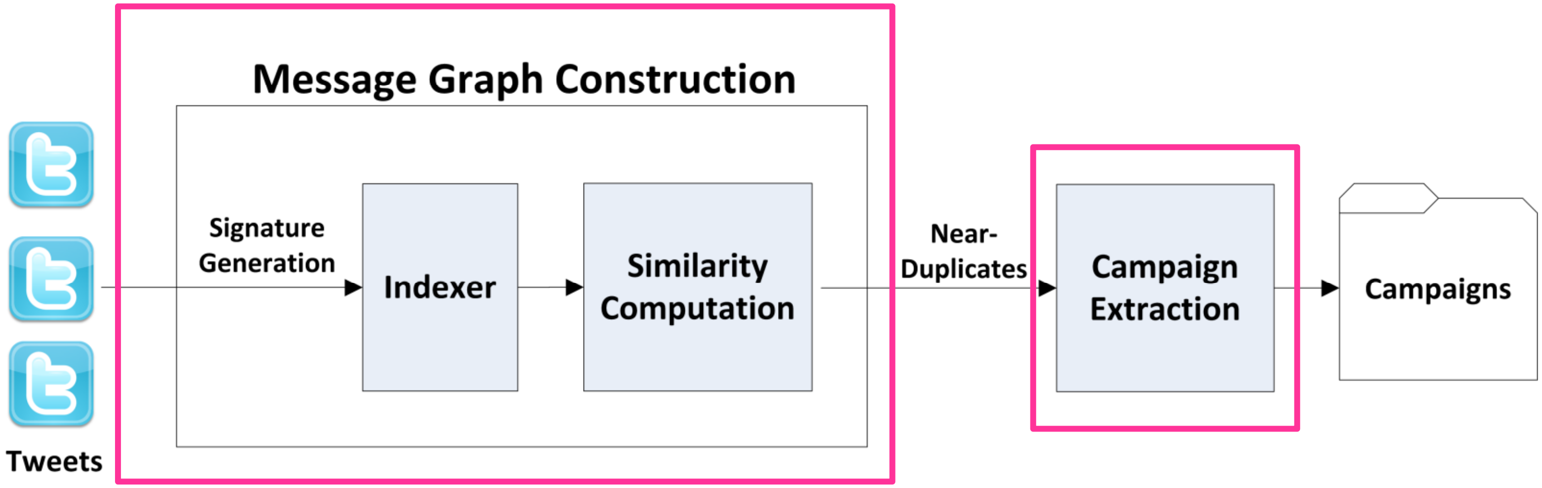
ID	Messages
1	Support Breast Cancer Awareness, add a #twibbon to your avatar now! - http://bit.ly/4DQ6vq
2	Support Breast Cancer Awareness, add a #twibbon to your avatar now! - http://bit.ly/3mAWR1
3	I'm having fun with @formspring. Create an account and follow me at http://formspring.me/xnadjeaaa
4	@Wookiefoot Real Money Doubling Forex Robot Fap Turbo 129$ http://bit.ly/ch9r1Hn?=mjkx
5	@justinbebier Support Breast Cancer Awareness, add a #twibbon to your avatar now! - http://bit.ly/4DQ6vq
6	RT @justinbebier Support … #twibbon to your avatar now! - http://bit.ly/4DQ6vq





Lee, K., Caverlee, J., Cheng, Z., and Sui, D. Campaign Extraction from Social Media. In *ACM TIST, Vol. 5, No. 1*, Dec. 2013

Two Key Components



- Message Graph Construction
  - Node: a message, Edge: if a pair of messages (nodes) are similar, add an edge
  - Measure message similarity by near-duplicate detection algorithm
  - Use MapReduce framework to improve efficiency

- Campaign (subgraph) Extraction
  - Find subgraphs each of which is dense like maximal clique
  - Use effective and efficient algorithm for campaign extraction

- Twitter Datasets (Short Text)
  - Small dataset – 1,912 messages
  - Large dataset – 1.5 million messages

Message Graph Construction

- Identifying correlated messages for Message Graph Construction
  - Unigram
  - Shingling
  - I-Match
  - SpotSigs

Message = “i think lady gaga is unique person”

4-**Shingling**: {“i think lady gaga”, “think lady gaga is”, “lady gaga is unique”, “gaga is unique person”}

I-Match: {“think”, “lady”, “gaga”, “unique”, “person”} → {“gaga”, “lady”, “person”, “think”, “unique”} -> {“gagaladypersonthinkunique”}

SpotSigs: {“i:lady:gaga”, “think:lady:gaga”, “is:unique:person”}

$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Identifying Correlated Messages

- 1,912 messages (know ground truth)
  - 298 pairs of similar messages

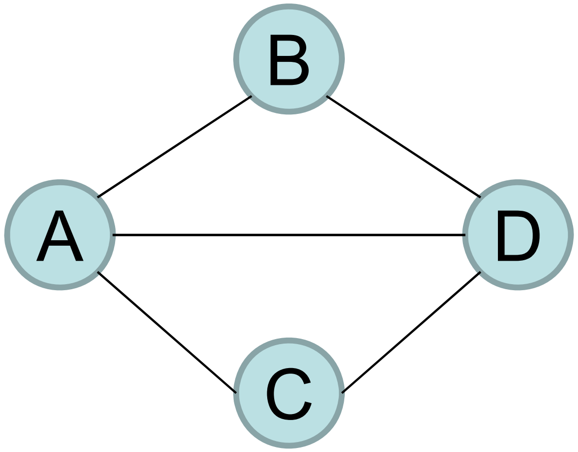
- Experimental results for Identifying correlated messages

Approach	$\mathbf{F}_1$	**Precision**	**Recall**
Unigram (τ = 0.8)	0.63	0.97	0.46
4-Shingling (τ = 0.3)	**0.81**	0.89	0.73
I-Match (IDF=[0.0, 0.8])	0.50	0.53	0.47
SpotSigs (#A=500, τ = 0.4)	0.70	0.77	0.64

Campaign (subgraph) Extraction

- K-means clustering algorithm
- Loose campaign extraction (maximally connected components)
- Strict campaign extraction (maximal cliques)
- Cohesive campaign extraction (approximate approach to extract densely connected components)

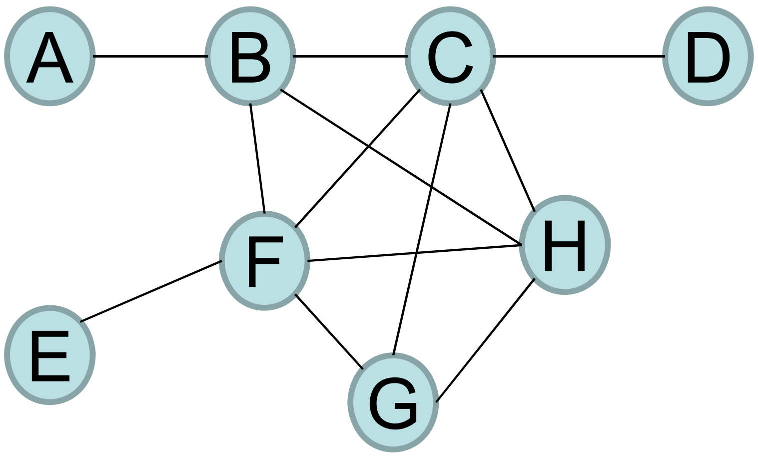
Cohesive Campaign Extraction

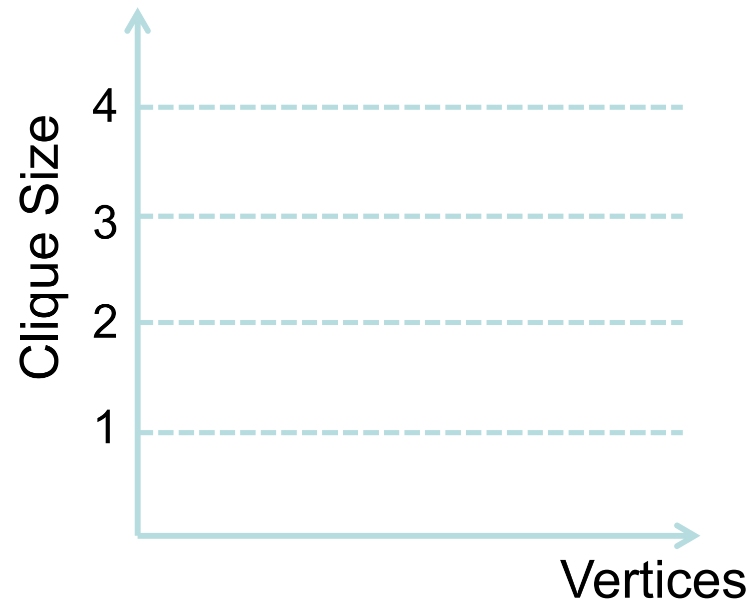


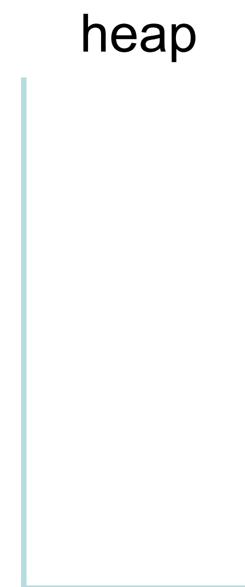
- Maximum co-clique size CC(x,y):
  - The biggest clique in the graph such that both vertices are members of the clique
  - CC(A,B) = 3
- Maximum clique size C(x):
  - The biggest clique it can participate
  - C(A) = 4

Wang et al. CSV: visualizing and mining cohesive subgraphs. In SIGMOD, 2008.

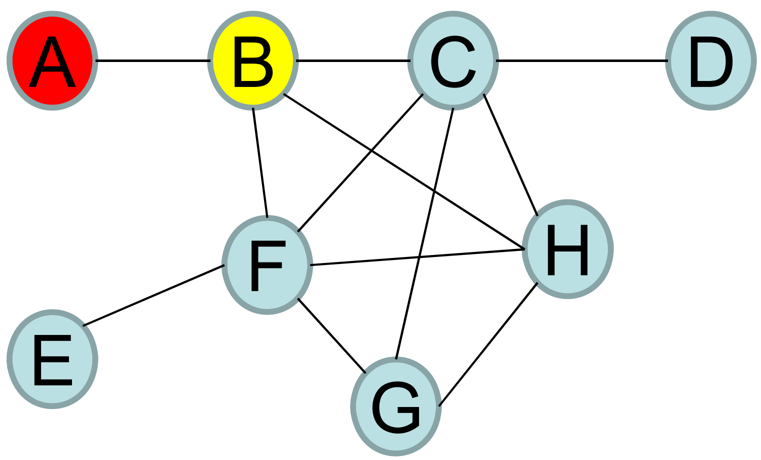
Cohesive Campaign Extraction

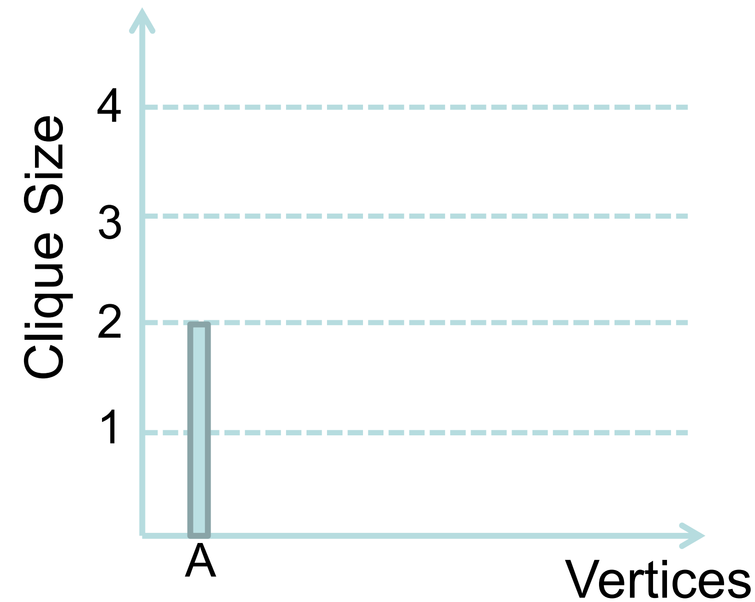


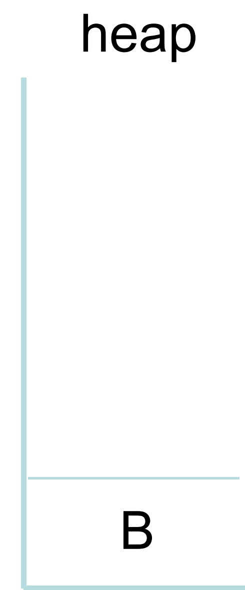




Cohesive Campaign Extraction

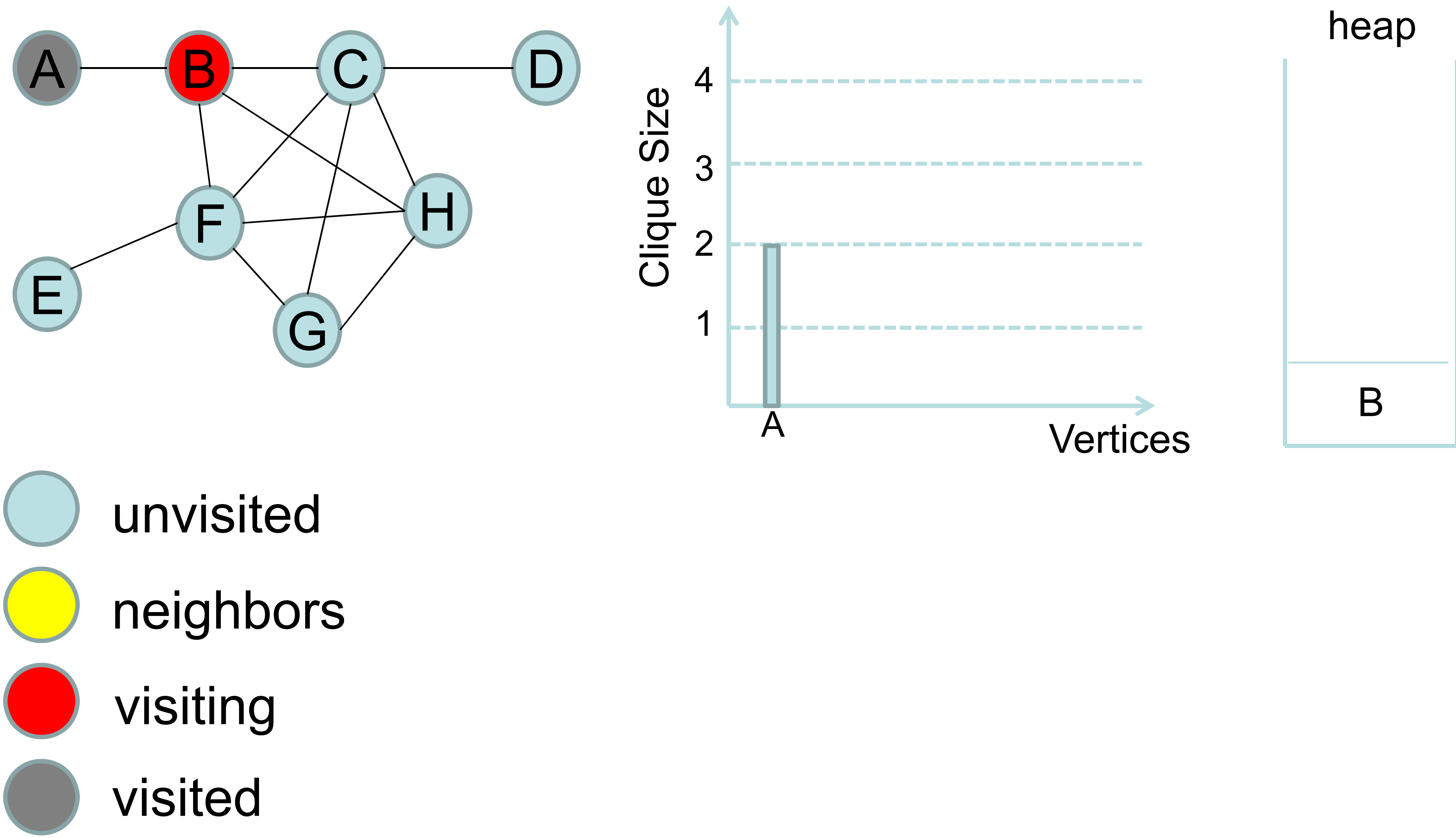




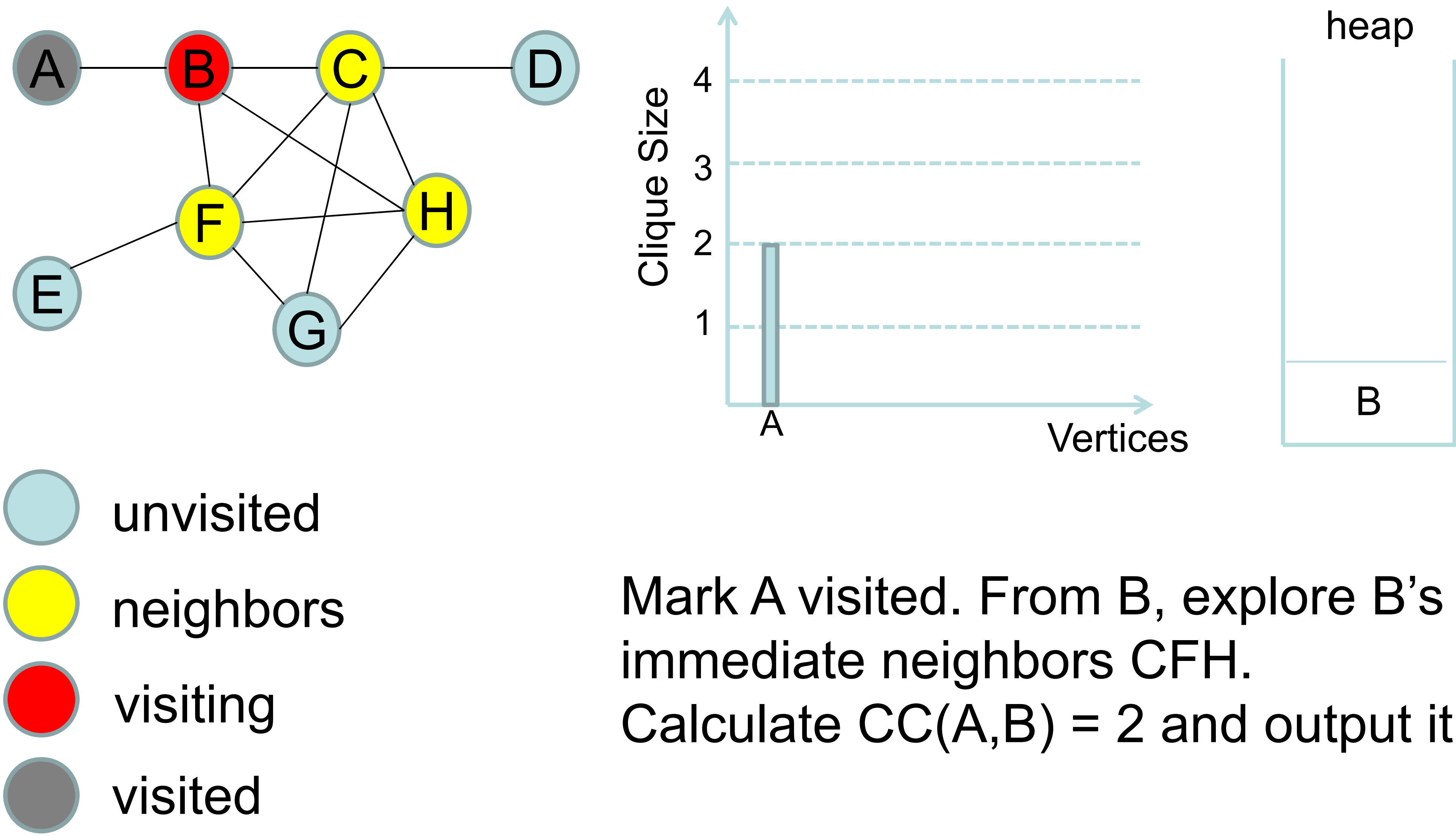


Start from A, explore A's neighbor B.
Calculate C(a) = 2 and output it.

Cohesive Campaign Extraction

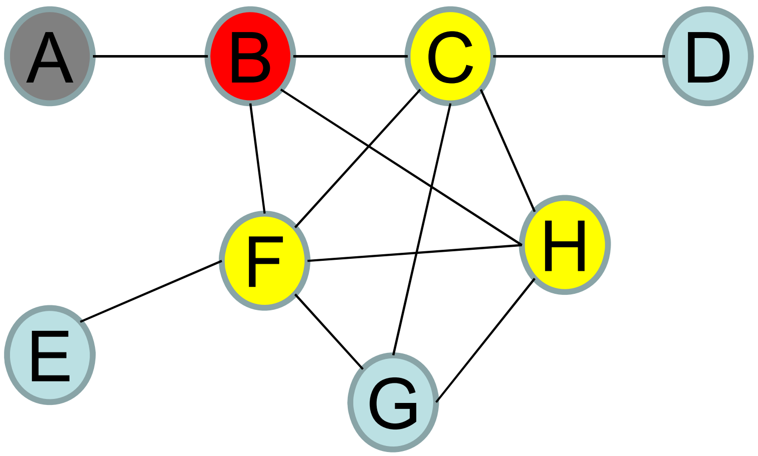


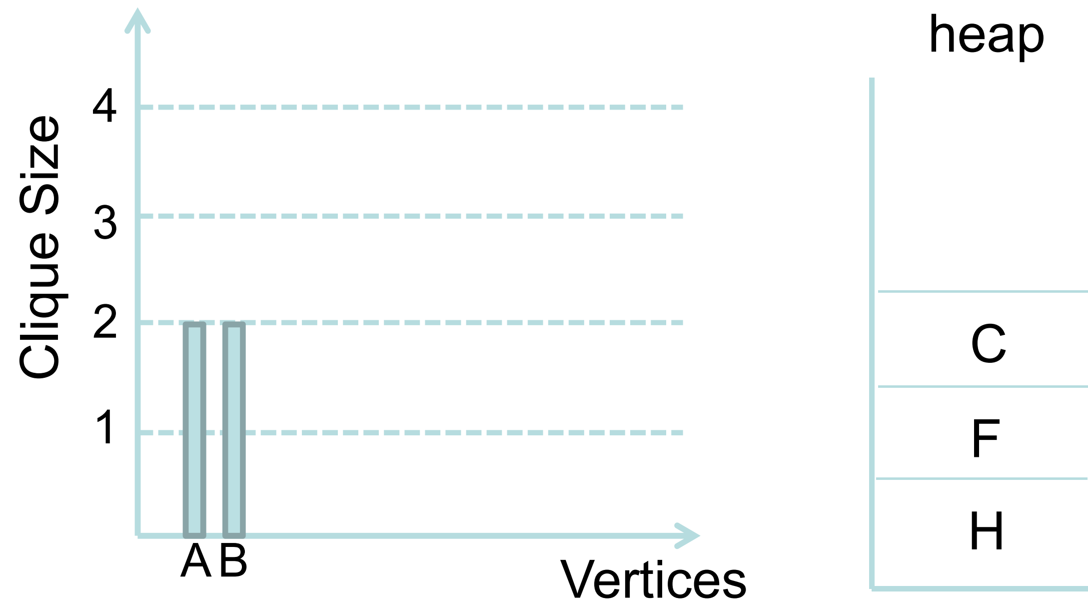
Cohesive Campaign Extraction



Mark A visited. From B, explore B's immediate neighbors CFH. Calculate CC(A,B) = 2 and output it.

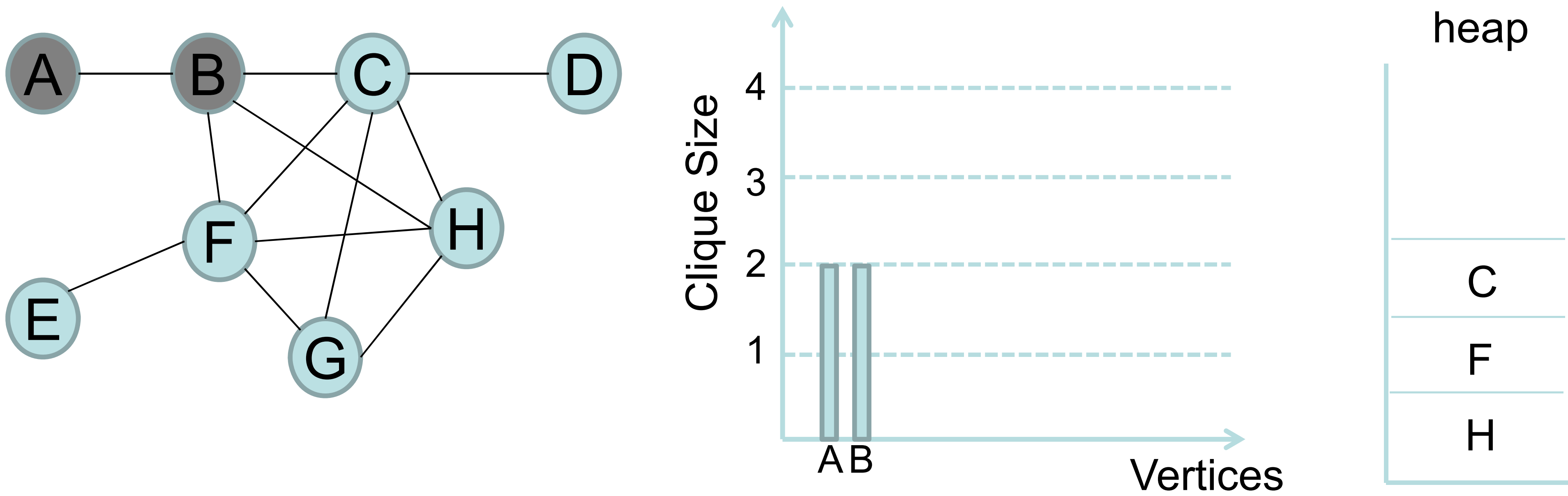
Cohesive Campaign Extraction



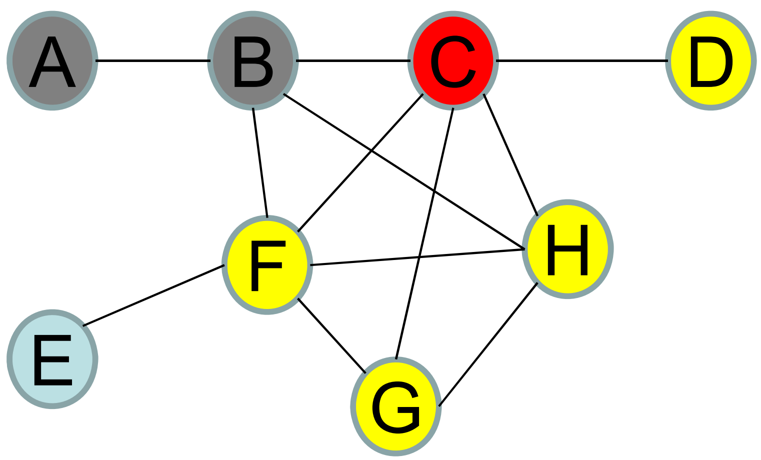


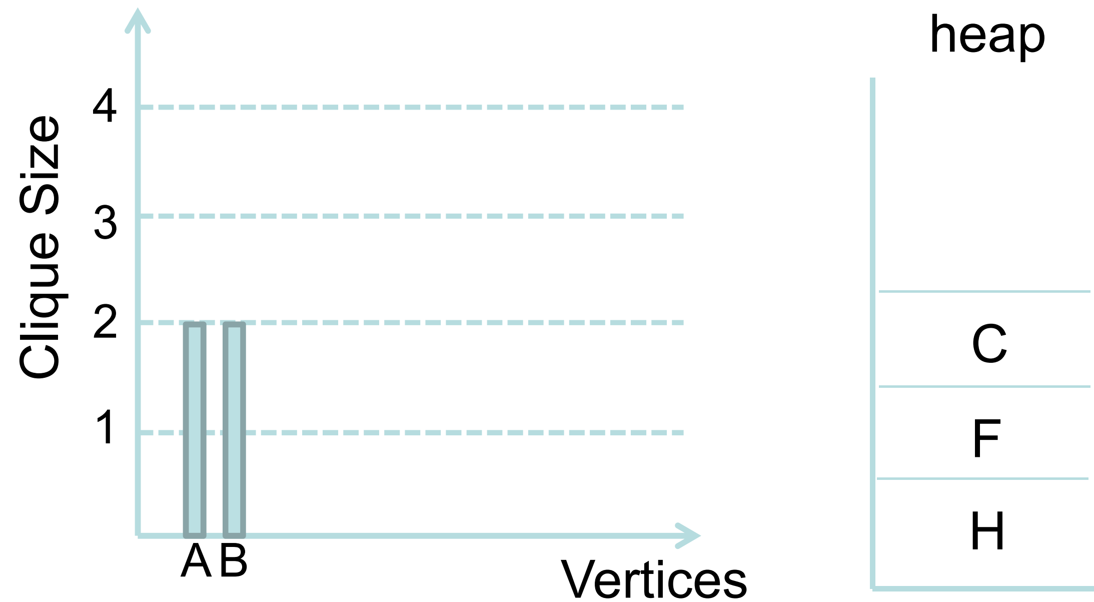
Mark A visited. From B, explore B's immediate neighbors CFH. Calculate CC(A,B) = 2 and output it.

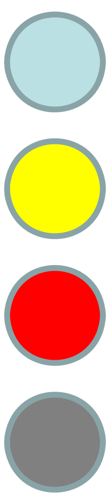
Cohesive Campaign Extraction



Cohesive Campaign Extraction







unvisited

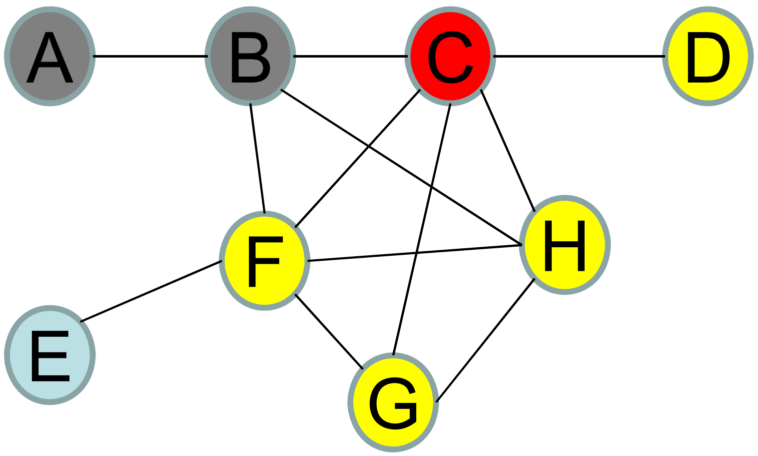
neighbors

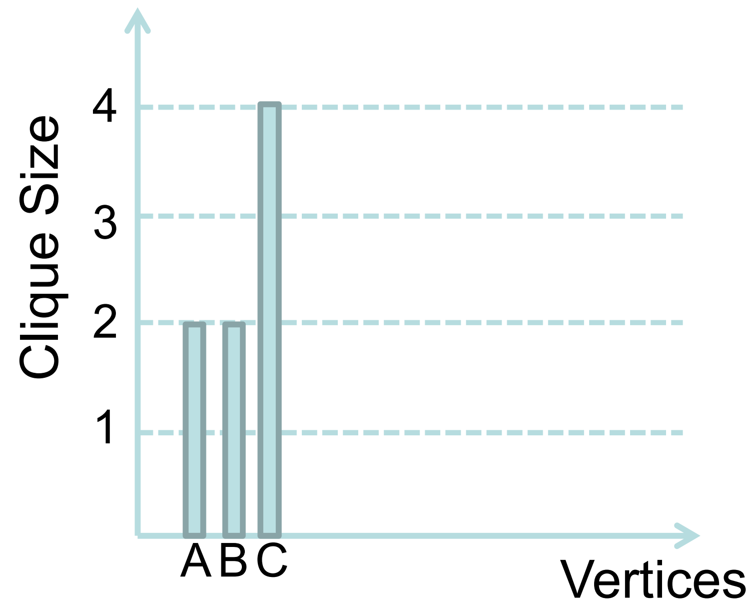
visiting

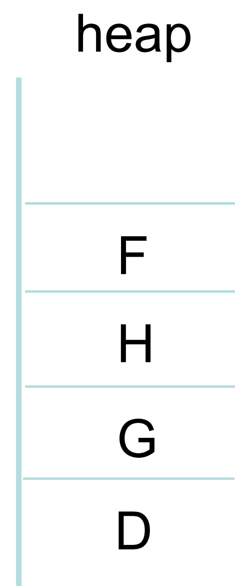
visited

Mark B visited. Choose C as next visiting vertex. From C, explore C's immediate neighbors DFGH. Calculate CC(B,C) = 4 and output it.

Cohesive Campaign Extraction

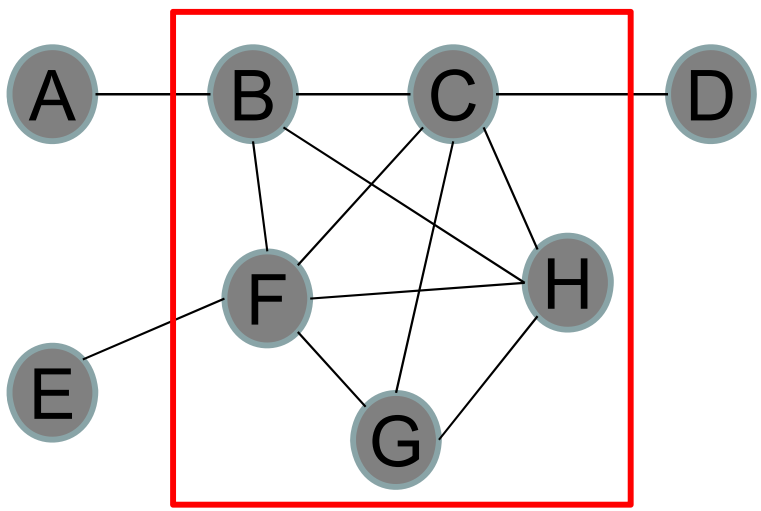


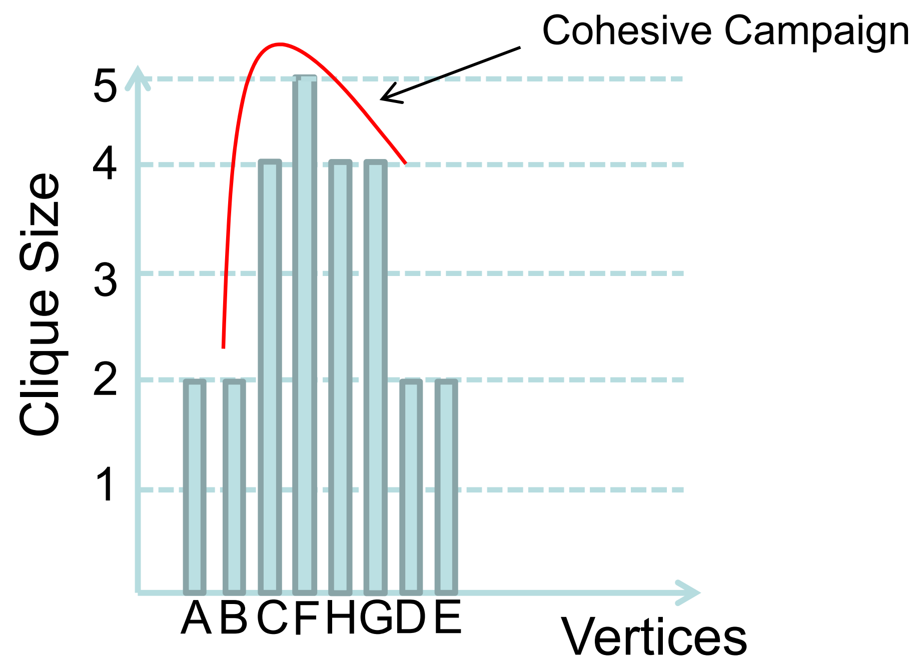




Mark B visited. Choose C as next visiting vertext.
From C, explore C’s immediate neighbors DFGH.
Calculate CC(B,C) = 4 and output it.

Cohesive Campaign Extraction





Visit every vertex accordingly.

The curve represents a cohesive campaign.

Campaign (subgraph) Extraction

- 1,912 messages (know ground truth)
  - 298 pairs of similar messages
  - 11 true campaigns
- Effectiveness Comparison of Campaign Detection Approaches

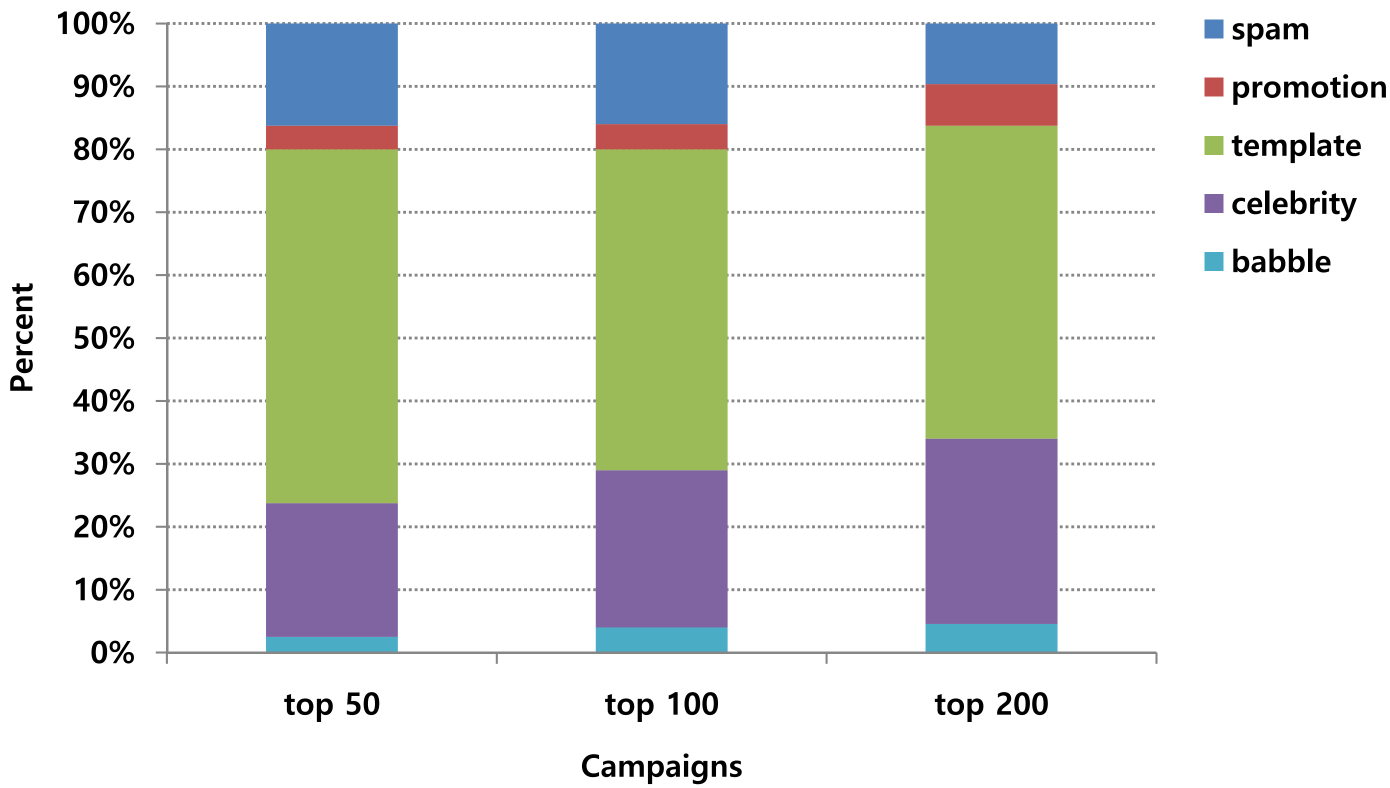
Approach	**NumC**	$\mathbf{F_1}$	**Precision**	**Recall**
Loose	12	0.962	0.986	0.940
Strict	12	0.906	0.907	0.904
Cohesive	11	**0.963**	0.977	0.950
k-means	5	0.89	1	0.805

So Far…

- Looked at a smallish dataset (with ground truth).
- 4-shingling and cohesive campaign extraction are the best approaches for message graph construction and campaign extractions.
- Next, apply these approaches to “the wild”.

Campaigns in the Wild

- 1.5 million messages → 7,033 campaigns (>= 4 messages)
- Five campaign categories -- 200 campaigns (>= 32 messages)
  - Spam, promotion, template, celebrity and babble campaigns



Examples of Campaigns

Spam Campaigns

#Monthly Iron Man 2 (Three-Disc Blu-ray ... http://bit.ly/9L0aZU

#getit Iron Man 2 (Three-Disc Blu-ray ... http://bit.ly/bREezs

#FollowWednesday Iron Man 2 (Three-Disc Blu-ray ... http://bit.ly/9haKNB

@Judd6149 Did you know you can view … http://tinyurl.com/ch7d5b

@Gleneagleshotel Did you know you can view … http://tinyurl.com/ybtfzys

@Re_Reading Did you know you can view ... http://tinyurl.com/ybtfzys

Promotion Campaign

#FightPediatricCancer! RT and Dreyer's Fruit Bars will donate $1 …. http://bit.ly/aZudoJ

RT @SupportSPN: #FightPediatricCancer! RT and Dreyer's Fruit Bars will donate $1 … http://bit.ly/aZudoJ

#FightPediatricCancer! RT and Dreyer's Fruit Bars will donate $1 … http://bit.ly/aZudoJ via @zaibatsu

Template Campaign

I posted a new photo to Facebook http://fb.me/KDa8EtY8

I posted a new photo to Facebook http://fb.me/CnFXpQvc

I posted a new photo to Facebook http://fb.me/uwxJShsV

Celebrity Campaign

@justinbieber pleaseFollow me please

@justinbieber Please follow me I love you really!

@justinbieber please follow me :] i love you ♥

Babble Campaign

I'm so tired!

I'm so tired today

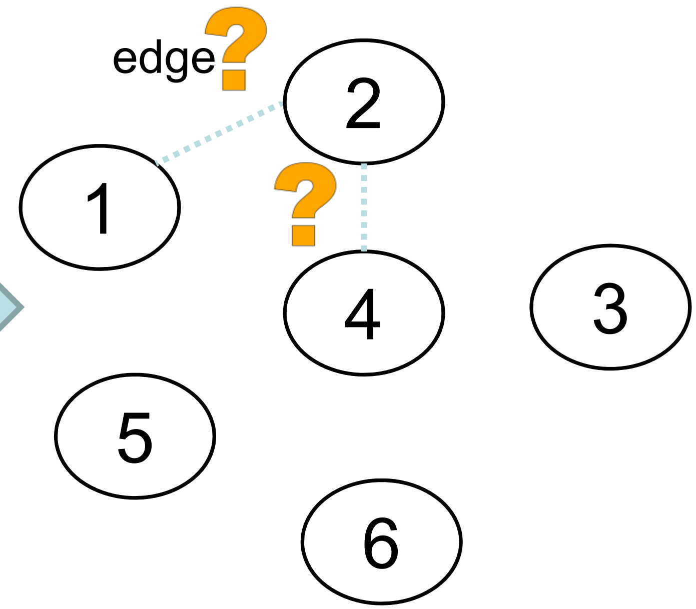
I'm so tired omg

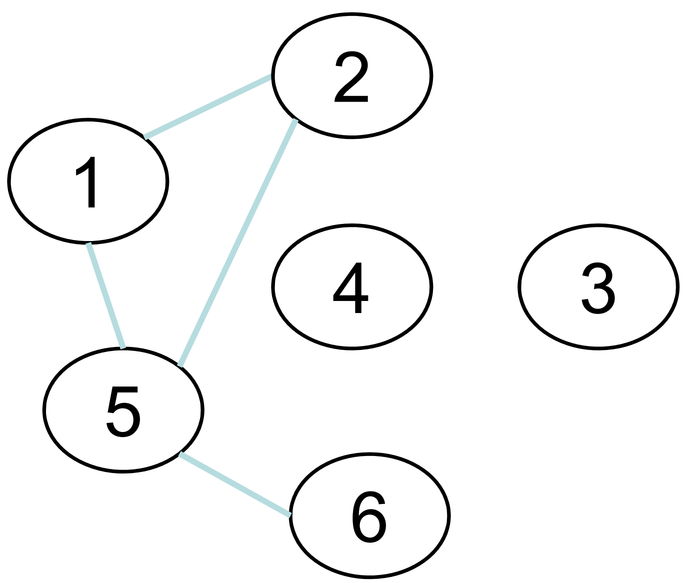
Top-10 Largest Campaigns

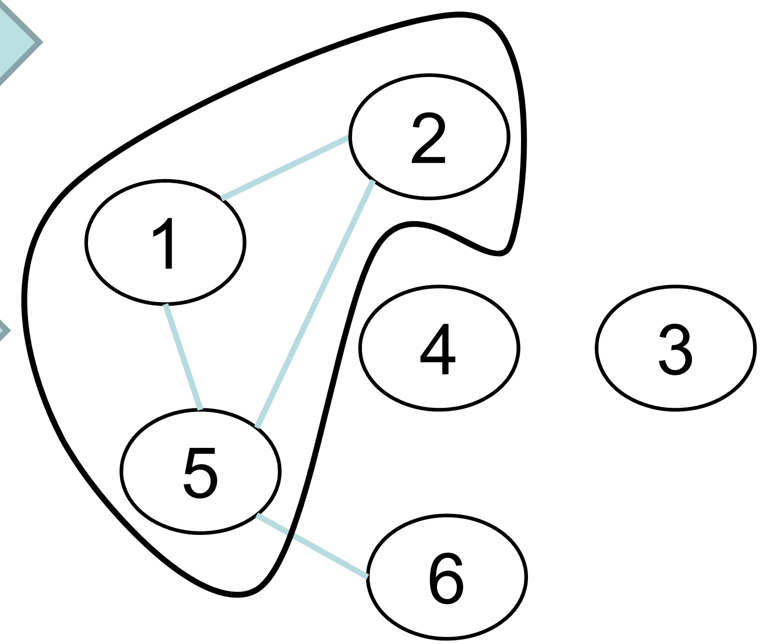
Msgs	Users	Talking Points
560	34	Iron Man 2 spam
401	390	Facebook photo template
231	231	Support Breast Cancer Research (short link)
218	218	Formspring template
203	197	Chat template (w/ link)
166	166	Support Breast Cancer Research (full link)
165	154	Quote "send to anyone u don't regret meeting"
153	153	Justin Bieber Retweets
145	31	Twilight Movie spam
111	111	Quote "This October has 5 Fridays ..."

User Level Campaign Detection

User ID	User Messages
1	M1: Support Breast Cancer Awareness, add a #twibbon M2: your avatar now! - http://bit.ly/4DQ6vq
2	M1: Support Breast Cancer Awareness, add a #twibbon M2: your avatar now! - http://bit.ly/3mAWR1
3	M1: I'm having fun with @formspring. Create an account M2: follow me at http://formspring.me/xnadjeaaa
4	M1: @Wookiefoot Real Money Doubling Forex Robot Fap M2: Turbo 129$ http://bit.ly/ch9r1Hn?=mjkx
5	M1: @justinbebier Support Breast Cancer Awareness, add M2: your avatar now! - http://bit.ly/4DQ6vq
6	M1: RT @justinbebier Support … #twibbon to M2: your avatar now! - http://bit.ly/4DQ6vq

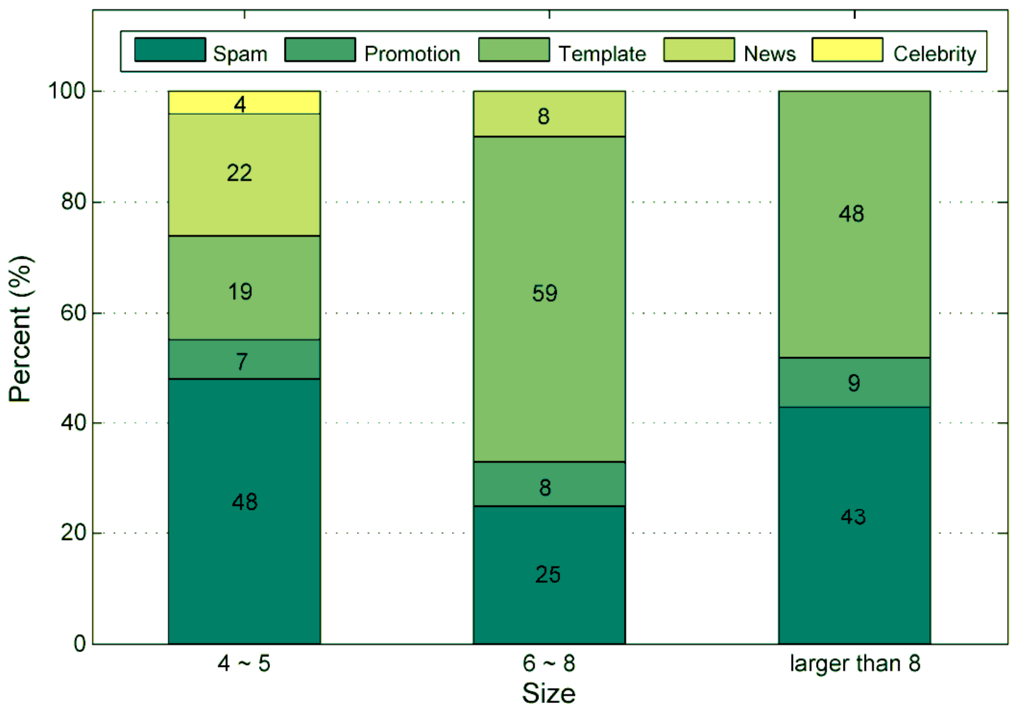






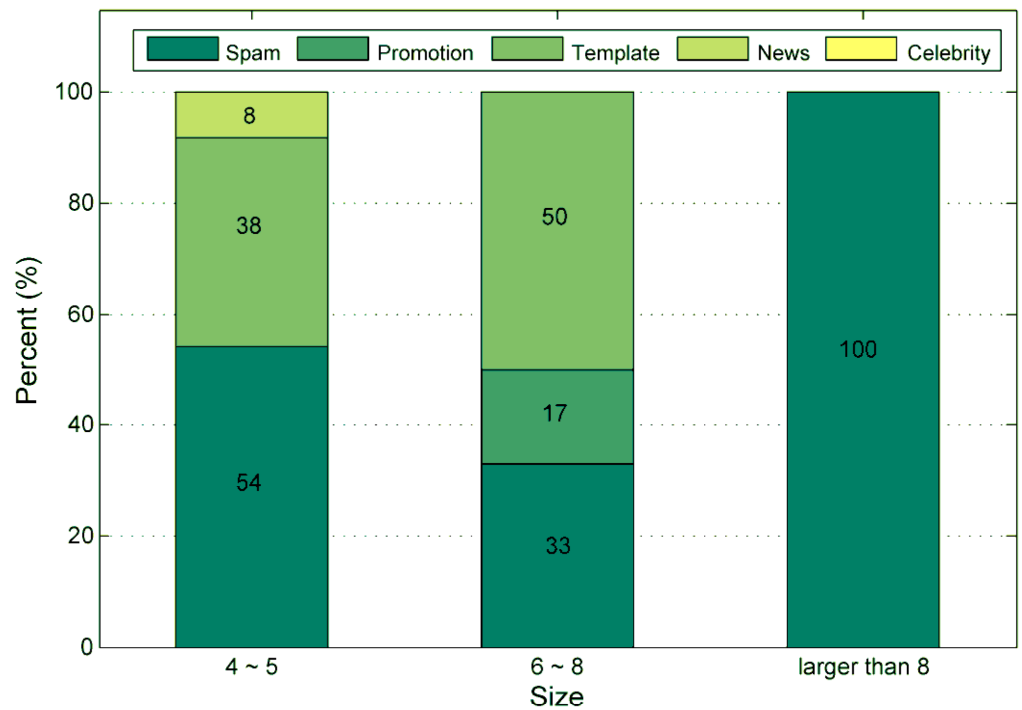
User Level Campaign Detection

62 campaigns (>= 4 users)



Campaign Type Distribution (threshold: 20% similarity)

28 campaigns (>= 4 users)



Campaign Type Distribution (threshold: 50% similarity)

The higher threshold is, the larger the proportion of inorganic campaigns is.

So far… Campaign Detection Approaches

- Graph-based spam campaign detection

- Content-driven campaign detection

Reference List

- Gao, H., Hu J., Wilson, C., Li, Z., Chen, Y., and Zhao, B. Detecting and characterizing social spam campaigns. In IMC, 2010.
- Lee, K., Caverlee, J., Cheng, Z., and Sui, D. Content-Driven Detection of Campaigns in Social Media. In CIKM, 2011
- Lee, K., Caverlee, J., Cheng, Z., and Sui, D. Campaign Extraction from Social Media. In ACM TIST, Vol. 5, No. 1, December 2013.
- Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Flammini, A., and Menczer, F. Detecting and Tracking Political Abuse in Social Media. In ICWSM, 2011.
- Mukherjee, A., Liu, B., and Glance, N. Spotting fake reviewer groups in consumer reviews. In WWW, 2012.

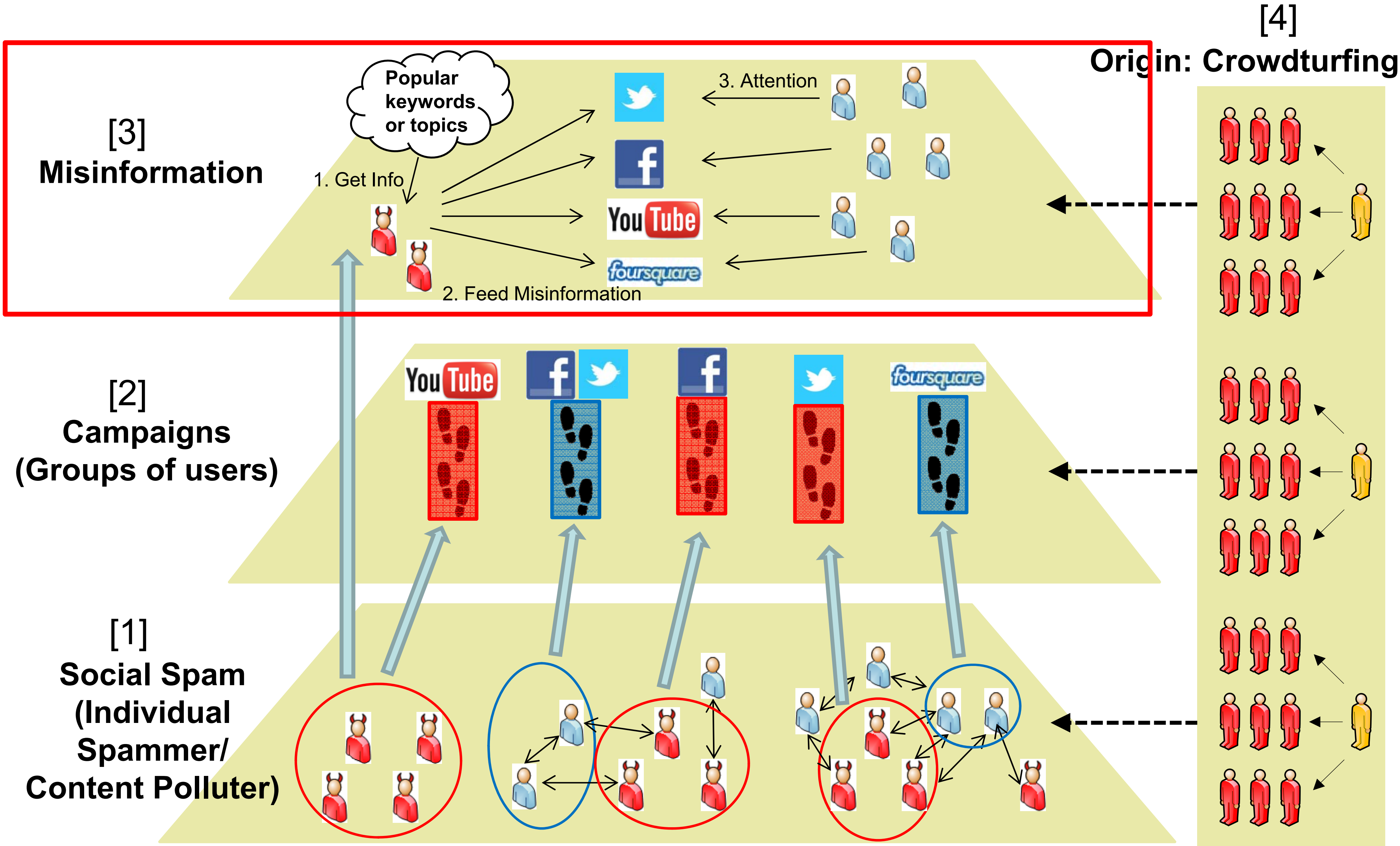
Schedule

14:00 ~ 14:10	Introduction to Social Media Threats (Social Spam, Campaigns, Misinformation and Crowdturfing)
14:10 ~ 14:55	Social Spam
14:55 ~ 15:30	Campaigns
15:30 ~ 16:00	Break
16:00 ~ 16:30	Misinformation
16:30 ~ 17:10	Crowdturfing
17:10 ~ 17:30	Challenges, Opportunities and Conclusion

Schedule

14:00 ~ 14:10	Introduction to Social Media Threats (Social Spam, Campaigns, Misinformation and Crowdturfing)
14:10 ~ 14:55	Social Spam
14:55 ~ 15:30	Campaigns
15:30 ~ 16:00	Break
16:00 ~ 16:30	Misinformation
16:30 ~ 17:10	Crowdturfing
17:10 ~ 17:30	Challenges, Tools and Conclusion

Conceptual Level of Tutorial Theme



Misinformation Detection Approach

- Supervised misinformation detection approach
  - Detecting false news events on Twitter
  - Detecting fake images on Twitter during Hurricane Sandy

Detecting false news events on Twitter

Chileans love Twitter

- Prominent role for communications
  - online and offline
- All public figures tweet
- Well integrated with traditional media
  - E.g., Earthquake in Feb 27, 2010.

Castillo, C., Mendoza, M., and Poblete, B. Information credibility on twitter. In WWW, 2011.D

Twitter helped, but ...

- Large majority of tweets were very helpful

- Some tweets were not
  - False tsunami warnings
  - False reports of looting
  - ...

Table 4: Classification results for cases studied of *confirmed truths* and *false rumors*.

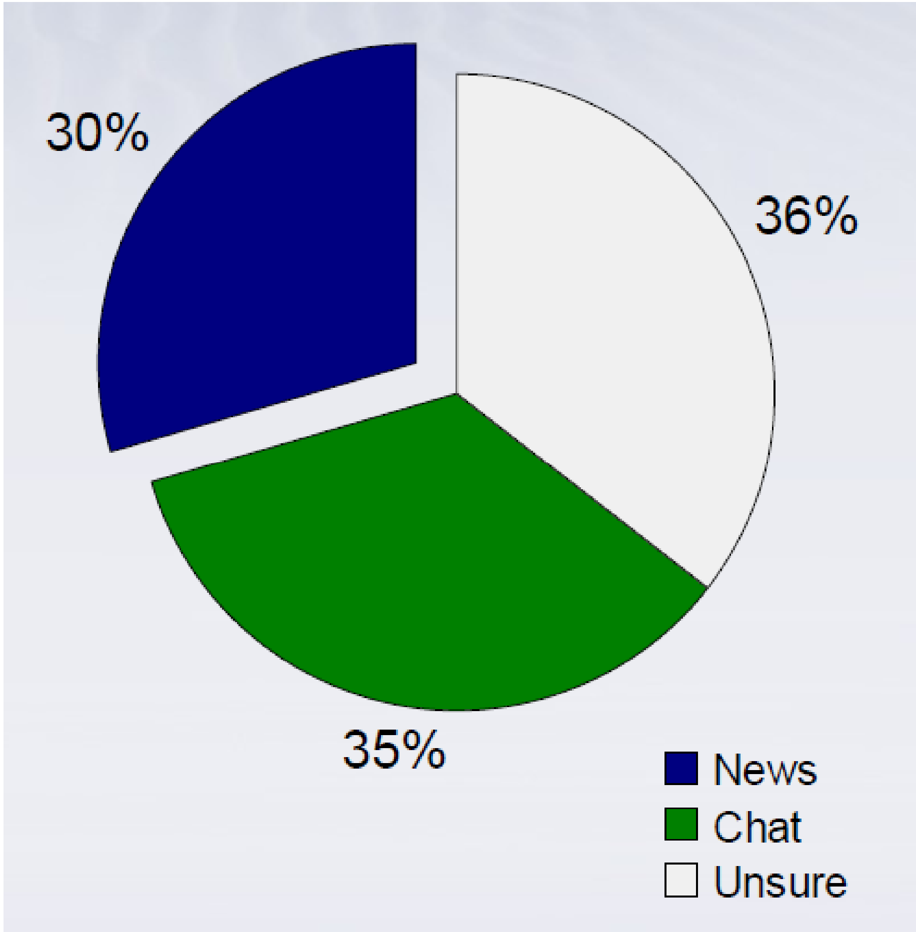
Case	# of unique tweets	% of re-tweets	# of unique "affirms"	# of unique "denies"	# of unique "questions"
Confirmed truths					
The international airport of Santiago is closed	301	81	291	0	7
The *Viña del Mar International Song Festival* is canceled	261	57	256	0	3
Fire in the Chemistry Faculty at the University of Concepción	42	49	38	0	4
Navy acknowledges mistake informing about tsunami warning	135	30	124	4	6
Small aircraft with six people crashes near Concepción	129	82	125	0	4
Looting of supermarket in Concepción	160	44	149	0	2
Tsunami in Iloca and Duao towns	153	32	140	0	4
TOTAL	1181		1123	4	30
AVERAGE	168,71		160,43	0,57	4,29
False rumors					
Death of artist Ricardo Arjona	50	37	24	12	8
Tsunami warning in Valparaiso	700	4	45	605	27
Large water tower broken in Rancagua	126	43	62	38	20
Cousin of football player Gary Medel is a victim	94	4	44	34	2
Looting in some districts in Santiago	250	37	218	2	20
"Huascar" vessel missing in Talcahuano	234	36	54	66	63
Villarrica volcano has become active	228	21	55	79	76
TOTAL	1682		502	836	216
AVERAGE	240,29		71,71	119,43	30,86

Supervised classification

- Goal: detecting false news events (sets of tweets)
- Approach:
  - Events (tweet sets) from TwitterMonitor
    - [Mathioudakis & Koudas 2010]
  - Labels from Amazon's Mechanical Turk
    - Event types: news, chat or unsure
    - Given news events, label each one to either credible or not
  - Built decision trees for each task

Labeling: News or Chat

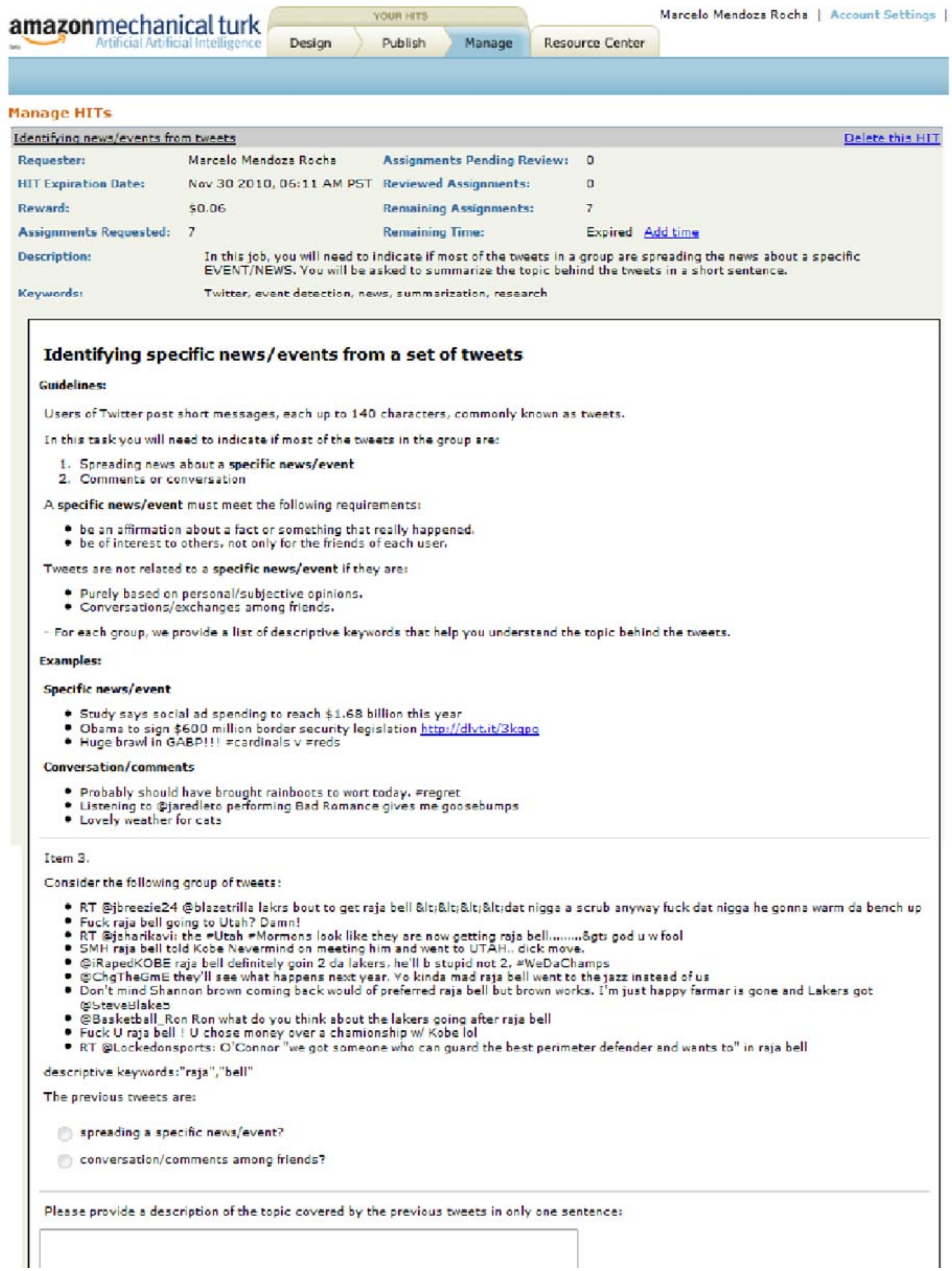
- 383 events from TwitterMonitor.net [Mathioudakis & Koudas]
- 7 evaluators per event
- >=5 agreement



Spreading a specific news/event

OR

Conversation or comments among friends.



amazon mechanical turk
Artificial Artificial Intelligence

YOUR HITS: Design | Publish | Manage | Resource Center

Marcelo Mendoza Rocha | Account Settings |

Manage HITs

Identifying news/events from tweets — Delete this HIT

Requester:	Marcelo Mendoza Rocha	Assignments Pending Review:	0
HIT Expiration Date:	Nov 30 2010, 06:11 AM PST	Reviewed Assignments:	0
Reward:	$0.06	Remaining Assignments:	7
Assignments Requested:	7	Remaining Time:	Expired Add time

Description: In this job, you will need to indicate if most of the tweets in a group are spreading the news about a specific EVENT/NEWS. You will be asked to summarize the topic behind the tweets in a short sentence.

Keywords: Twitter, event detection, news, summarization, research

Identifying specific news/events from a set of tweets

Guidelines:

Users of Twitter post short messages, each up to 140 characters, commonly known as tweets.

In this task you will need to indicate if most of the tweets in the group are:

1. Spreading news about a **specific news/event**
2. Comments or conversation

A **specific news/event** must meet the following requirements:

- be an affirmation about a fact or something that really happened.
- be of interest to others, not only for the friends of each user.

Tweets are not related to a **specific news/event** if they are:

- Purely based on personal/subjective opinions.
- Conversations/exchanges among friends.

- For each group, we provide a list of descriptive keywords that help you understand the topic behind the tweets.

Examples:

Specific news/event

- Study says social ad spending to reach $1.68 billion this year
- Obama to sign $600 million border security legislation http://dlvr.it/3kqpq
- Huge brawl in GABP!!! #cardinals v #reds

Conversation/comments

- Probably should have brought rainboots to wort today. #regret
- Listening to @jaredleto performing Bad Romance gives me goosebumps
- Lovely weather for cats

Item 3.

Consider the following group of tweets:

- RT @jbreezie24 @blazetrilla lakrs bout to get raja bell <<<<dat nigga a scrub anyway fuck dat nigga he gonna warm da bench up
- Fuck raja bell going to Utah? Damn!
- RT @jaharikavi: the #Utah #Mormons look like they are now getting raja bell.........> god u w fool
- SMH raja bell told Kobe Nevermind on meeting him and went to UTAH.. dick move.
- @iRapedKOBE raja bell definitely goin 2 da lakers, he'll b stupid not 2. #WeDaChamps
- @ChgTheGmE they'll see what happens next year. Yo kinda mad raja bell went to the jazz instead of us
- Don't mind Shannon brown coming back would of preferred raja bell but brown works. I'm just happy farmar is gone and Lakers got @SteveBlake5
- @Basketball_Ron Ron what do you think about the lakers going after raja bell
- Fuck U raja bell ! U chose money over a chamionship w/ Kobe lol
- RT @Lockedonsports: O'Connor "we got someone who can guard the best perimeter defender and wants to" in raja bell

descriptive keywords:"raja","bell"

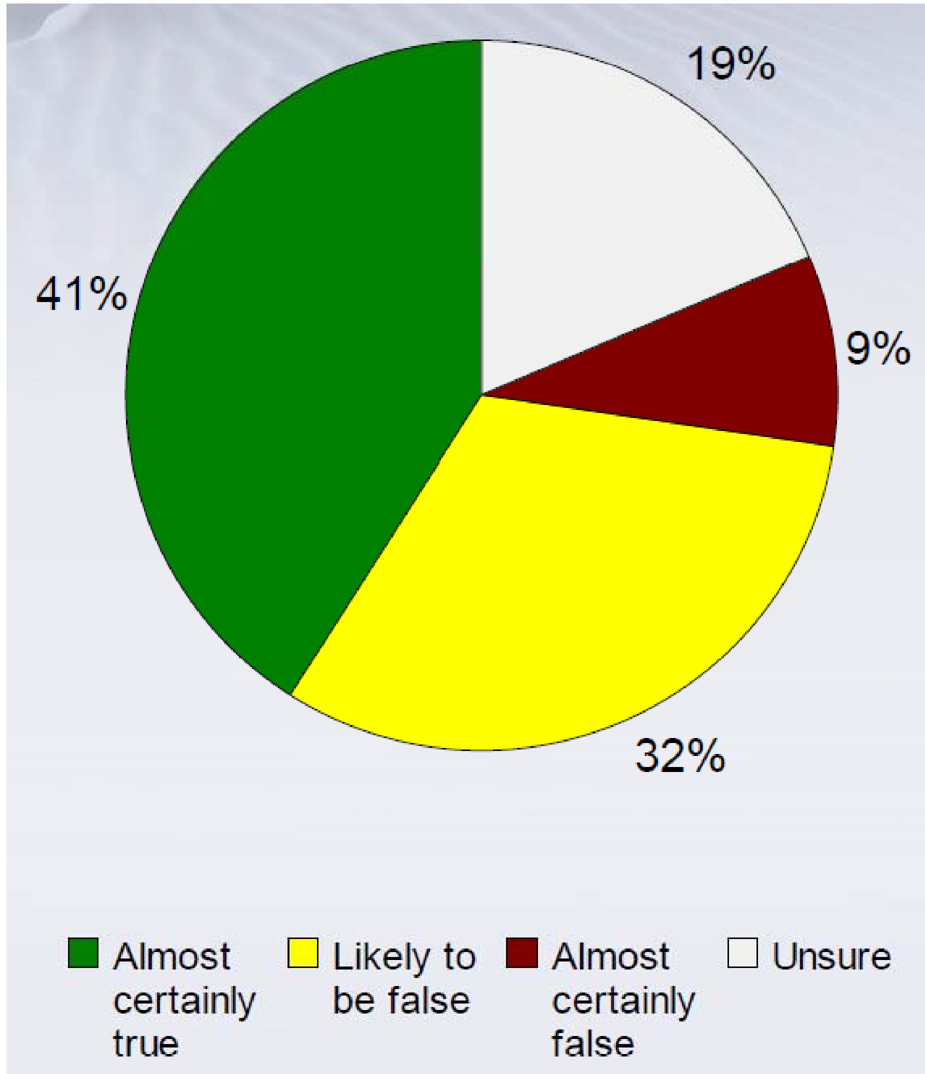
The previous tweets are:

- ○ spreading a specific news/event?
- ○ conversation/comments among friends?

Please provide a description of the topic covered by the previous tweets in only one sentence:

Labeling: Credible or Not

- 747 events automatically classified as news
- 7 evaluators per event
- >=5 agreement

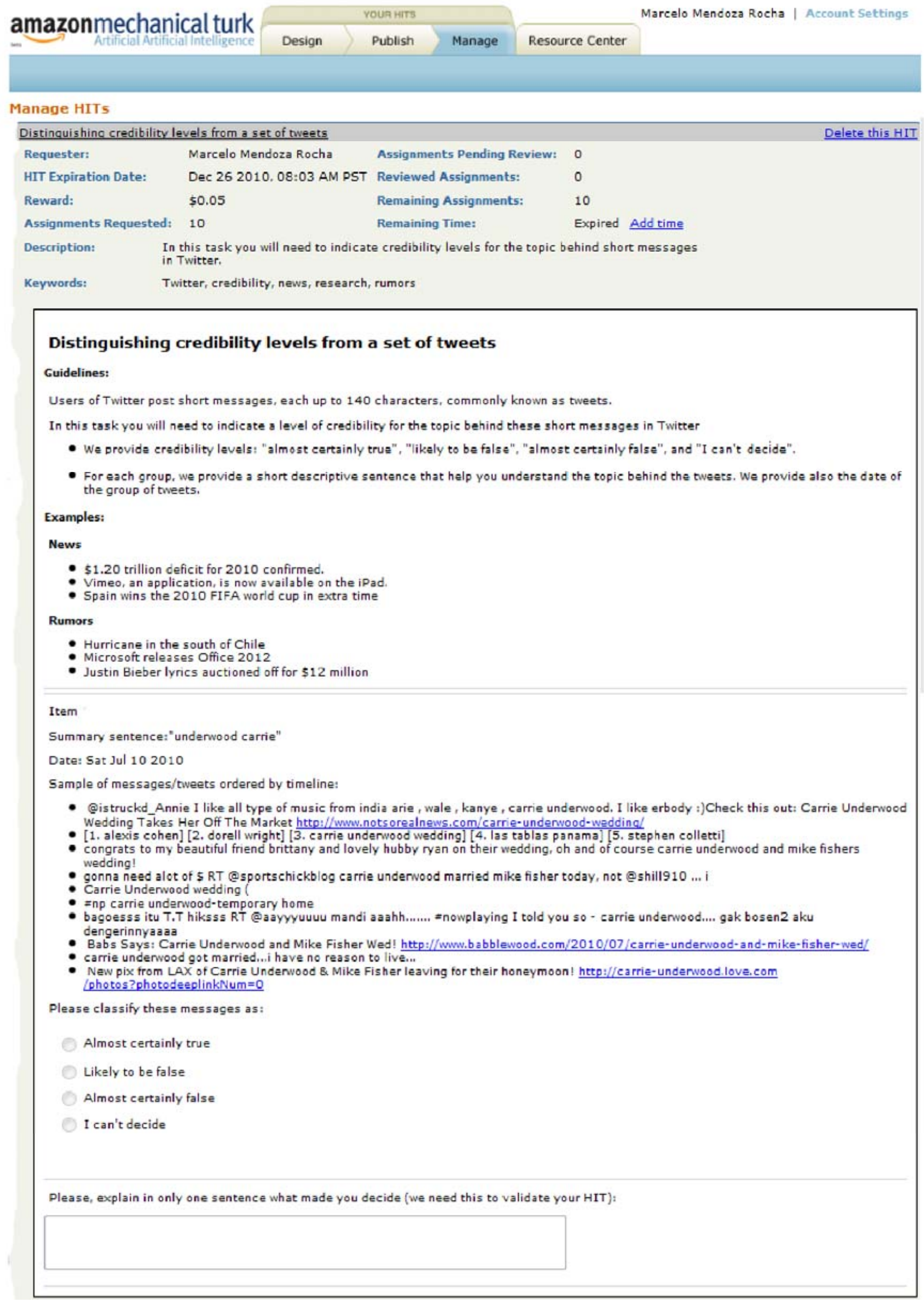


Almost certainly true

Likely to be true

Likely to be false

Almost certainly false


amazonmechanical turk
Artificial Artificial Intelligence

YOUR HITS | Design | Publish | Manage | Resource Center

Marcelo Mendoza Rocha | Account Settings

Manage HITs

Distinguishing credibility levels from a set of tweets — Delete this HIT

Requester:	Marcelo Mendoza Rocha	Assignments Pending Review:	0
HIT Expiration Date:	Dec 26 2010, 08:03 AM PST	Reviewed Assignments:	0
Reward:	$0.05	Remaining Assignments:	10
Assignments Requested:	10	Remaining Time:	Expired Add time

Description: In this task you will need to indicate credibility levels for the topic behind short messages in Twitter.

Keywords: Twitter, credibility, news, research, rumors

Distinguishing credibility levels from a set of tweets

Guidelines:

Users of Twitter post short messages, each up to 140 characters, commonly known as tweets.

In this task you will need to indicate a level of credibility for the topic behind these short messages in Twitter

- We provide credibility levels: "almost certainly true", "likely to be false", "almost certainly false", and "I can't decide".
- For each group, we provide a short descriptive sentence that help you understand the topic behind the tweets. We provide also the date of the group of tweets.

Examples:

News

- $1.20 trillion deficit for 2010 confirmed.
- Vimeo, an application, is now available on the iPad.
- Spain wins the 2010 FIFA world cup in extra time

Rumors

- Hurricane in the south of Chile
- Microsoft releases Office 2012
- Justin Bieber lyrics auctioned off for $12 million

Item

Summary sentence:"underwood carrie"

Date: Sat Jul 10 2010

Sample of messages/tweets ordered by timeline:

- @istruckd_Annie I like all type of music from india arie , wale , kanye , carrie underwood. I like erbody :)Check this out: Carrie Underwood Wedding Takes Her Off The Market http://www.notsorealnews.com/carrie-underwood-wedding/
- [1. alexis cohen] [2. dorell wright] [3. carrie underwood wedding] [4. las tablas panama] [5. stephen colletti]
- congrats to my beautiful friend brittany and lovely hubby ryan on their wedding, oh and of course carrie underwood and mike fishers wedding!
- gonna need alot of $ RT @sportschickblog carrie underwood married mike fisher today, not @shill910 ... i
- Carrie Underwood wedding (
- #np carrie underwood-temporary home
- bagoesss itu T.T hiksss RT @aayyyuuuu mandi aaahh...... #nowplaying I told you so - carrie underwood.... gak bosen2 aku dengerinnyaaaa
- Babs Says: Carrie Underwood and Mike Fisher Wed! http://www.babblewood.com/2010/07/carrie-underwood-and-mike-fisher-wed/
- carrie underwood got married...i have no reason to live...
- New pix from LAX of Carrie Underwood & Mike Fisher leaving for their honeymoon! http://carrie-underwood.love.com/photos?photodeeplinkNum=0

Please classify these messages as:

- Almost certainly true
- Likely to be false
- Almost certainly false
- I can't decide

Please, explain in only one sentence what made you decide (we need this to validate your HIT):

Credible tweets for users tend to ...

- Have a URL
- Don't have exclamation marks
- Express a negative sentiment
- Are re-posted by prolific users
- Are re-posted by well-connected users

Experimental Results

Table 4: Results for the classification of newsworthy topics.

Class	TP Rate	FP Rate	Prec.	Recall	F_1
NEWS	0.927	0.039	0.922	0.927	0.924
CHAT	0.874	0.054	0.892	0.874	0.883
UNSURE	0.873	0.07	0.86	0.873	0.866
W. Avg.	0.891	0.054	0.891	0.891	0.891

89% accuracy

Table 7: Results for the credibility classification.

Class	TP Rate	FP Rate	Prec.	Recall	F_1
A ("true")	0.825	0.108	0.874	0.825	0.849
B ("false")	0.892	0.175	0.849	0.892	0.87
W. Avg.	0.860	0.143	0.861	0.860	0.86

86% accuracy

Detecting fake images on Twitter during Hurricane Sandy

Background: Hurricane Sandy

- Dates: Oct 22 - 31, 2012
- Category 3 storm
- Damages worth $75 billion USD
- Coast of NE America [Atlantic ocean]

Gupta, A., Lamba, H., Kumaraguru, P., and Joshi, A. Faking Sandy: characterizing and identifying fake images on Twitter during Hurricane Sandy. In *WWW Companion*, 2013

Motivation

theguardian

Printing sponsored by:

Kodak
All-in-One Printers

USNEWS
BLOG

Hurricane Sandy brings storm of fake news and photos to New York

Misinformation over storm spread quickly online, abetted by journalists no longer taught importance of verifying every source

Motivation

Man faces fallout for spreading false Sandy reports on Twitter

By Doug Gross , CNN

October 31, 2012 -- Updated 2244 GMT (0644 HKT) | Filed under: Social Media

CNN.com

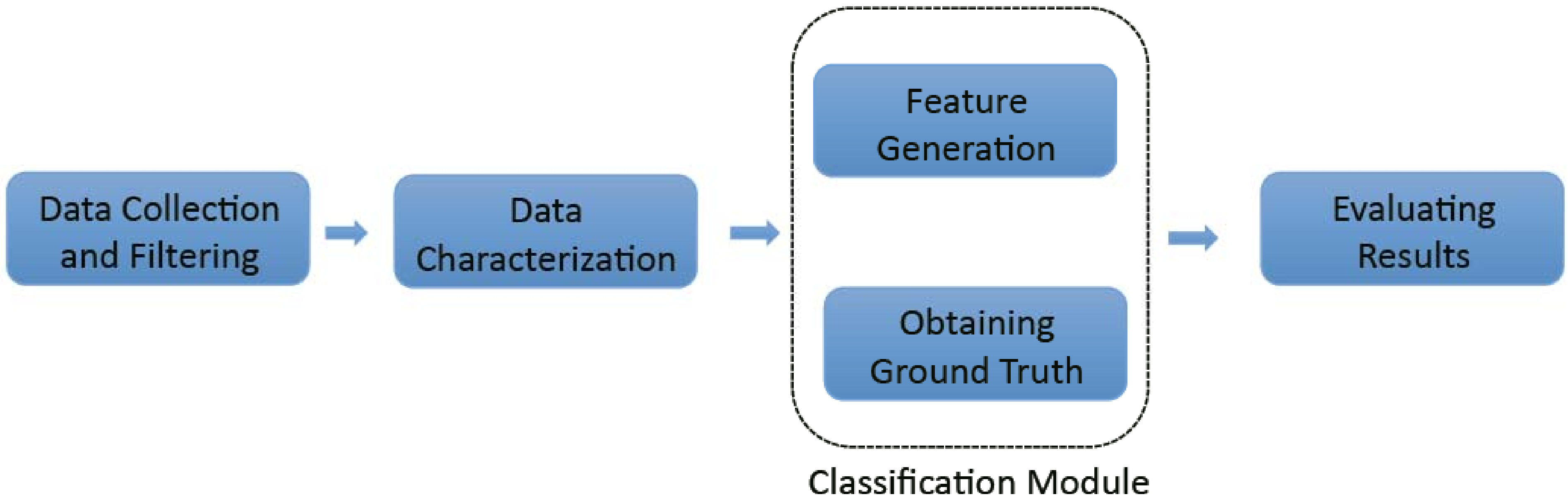
This tweet was one of several false reports posted by Twitter user @ComfortablySmug as Sandy pummeled New York.

(CNN) -- As Superstorm Sandy slammed into the East Coast on Monday night, one Twitter user in New York City posted a flurry of alarming reports about fallout from the storm -- from plans to shut down all power in Manhattan to floodwaters pouring into the New York Stock Exchange.

Like many social media messages about Sandy, they were scary and confusing, but some of them were reported as facts by news outlets.

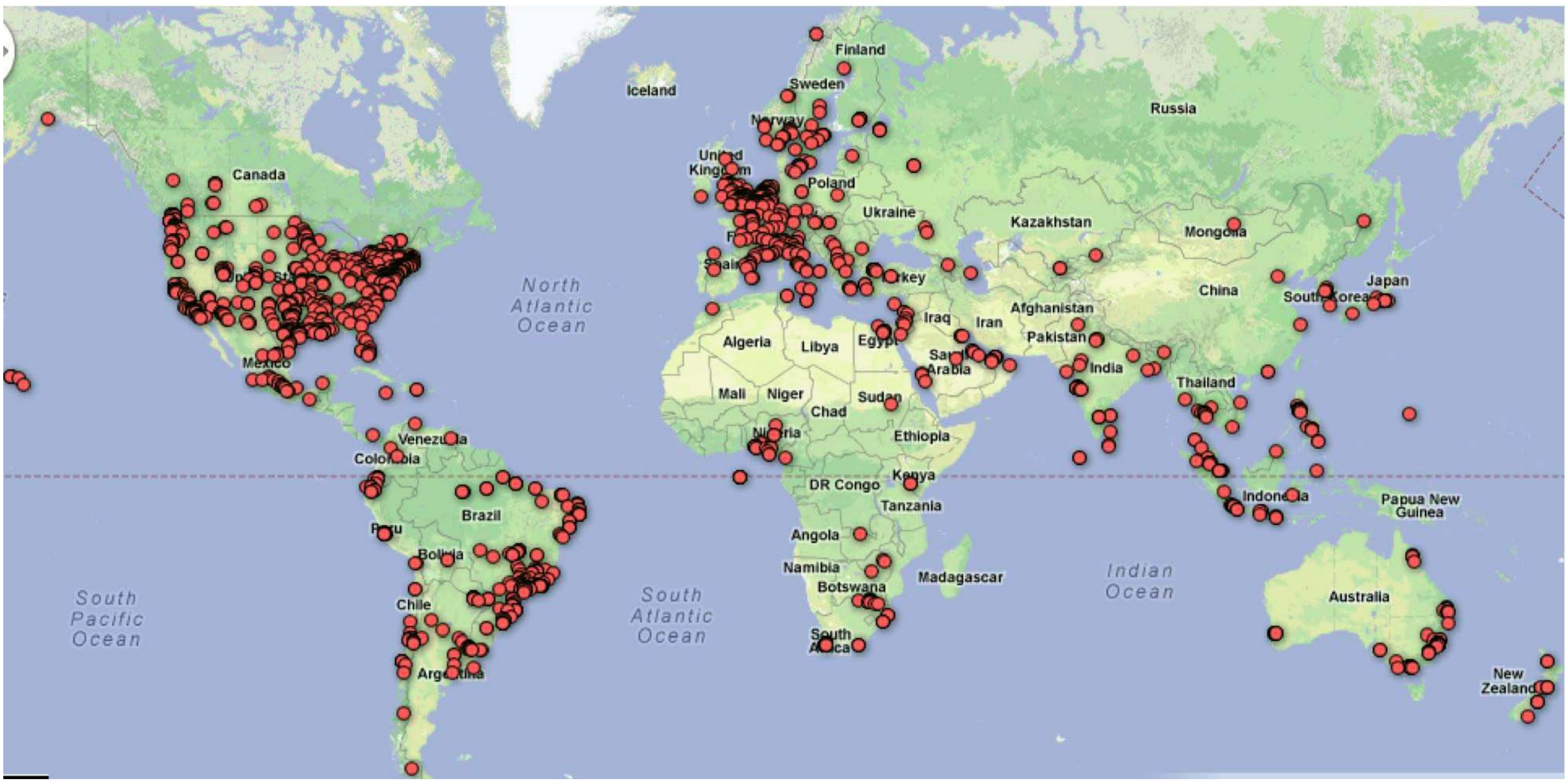
Goal and Methodology

- Goal: Detecting tweets containing fake images
- Methodology



Data Description – Total Sandy Dataset

Total Tweets	1,782,526
Total unique users	1,174,266
Tweets with URLs	622,860



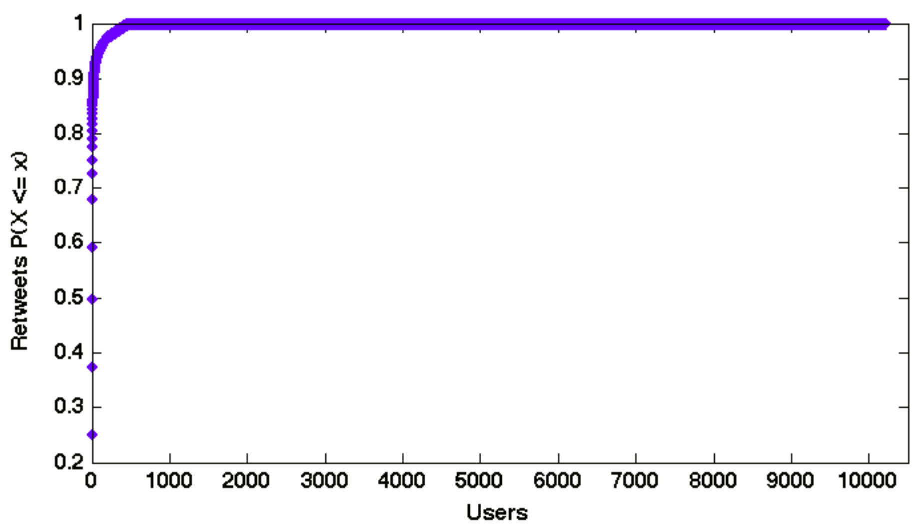
Data Filtering

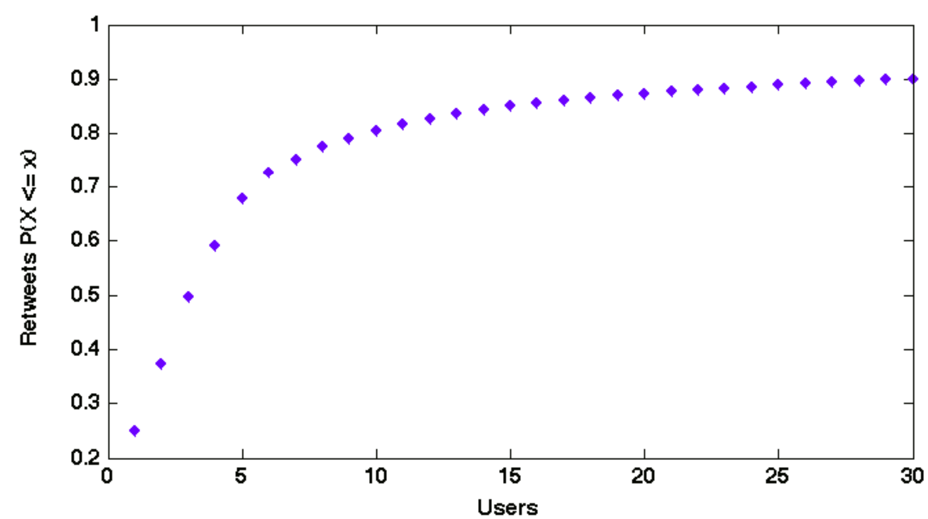
- Reputable online resource to filter fake and real images
  - Guardian collected and publically distributed a list of fake and true images shared during Hurricane Sandy

Tweets with fake images	10,350
Users with fake images	10,215
Tweets with real images	5,767
Users with real images	5,678

Characterization – Fake Image Propagation

- 86% of tweets spreading the fake images were retweets
- Top 30 users out of 10,215 users (0.3%) resulted in 90% of the retweets of fake images





Role of Explicit Twitter Network

- Crawled the Twitter network for all users who tweeted the fake image URLs

- Analyzed role of follower network in fake image propagation
  – Just 11% overlap between the retweet and follower graphs of tweets containing fake images

Classification

- 5 fold cross validation
- Randomly selected fake tweets equal to number of real tweets to prevent bias in the classification

User Features [F1]
Number of Friends
Number of Followers
Follower-Friend Ratio
Number of times listed
User has a URL
User is a verified user
Age of user account

Tweet Features [F2]
Length of Tweet
Number of Words
Contains Question Mark?
Contains Exclamation Mark?
Number of Question Marks
Number of Exclamation Marks
Contains Happy Emoticon
Contains Sad Emoticon
Contains First Order Pronoun
Contains Second Order Pronoun
Contains Third Order Pronoun
Number of uppercase characters
Number of negative sentiment words
Number of positive sentiment words
Number of mentions
Number of hashtags
Number of URLs
Retweet count

Classification Results

	F1 (user)	F2 (tweet)	F1+F2
Naïve Bayes	56.32%	91.97%	91.52%
Decision Tree	53.24%	97.65%	96.65%

- Best results were obtained from Decision Tree classifier, the researchers got 97% accuracy in predicting fake images from real.
- Tweet based features are very effective in distinguishing fake images tweets from real.

So far… Misinformation Detection Approach

- Supervised misinformation detection approach
  - Detecting false news events on Twitter
  - Detecting fake images on Twitter during Hurricane Sandy

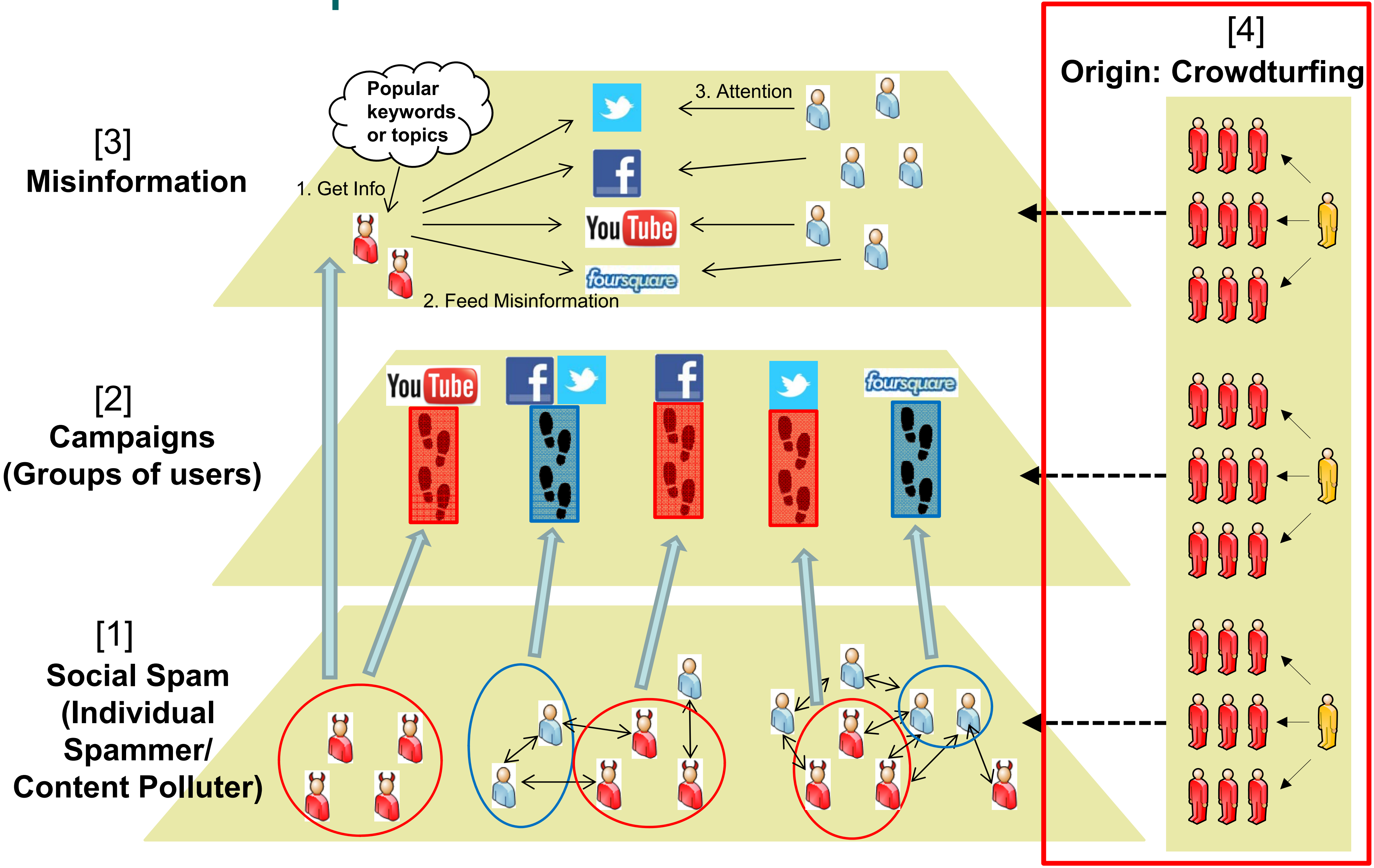
Reference List

- Castillo, C., Mendoza, M., and Poblete, B. Information credibility on twitter. In WWW, 2011.
- Yang, F., Liu, Y., Yu, X., and Yang, M. Automatic detection of rumor on Sina Weibo. In SIGKDD Workshop on Mining Data Semantics, 2012.
- Gupta, A., Lamba, H., Kumaraguru, P., and Joshi, A. Faking Sandy: characterizing and identifying fake images on Twitter during Hurricane Sandy. In WWW Companion, 2013.
- Xia, X., Yang, X., Wu, C., Li, S., and Bao, L. Information credibility on twitter in emergency situation. In Proceedings of the 2012 Pacific Asia conference on Intelligence and Security Informatics (PAISI), 2012.

Schedule

14:00 ~ 14:10	Introduction to Social Media Threats (Social Spam, Campaigns, Misinformation and Crowdturfing)
14:10 ~ 14:55	Social Spam
14:55 ~ 15:30	Campaigns
15:30 ~ 16:00	Break
16:00 ~ 16:30	Misinformation
16:30 ~ 17:10	Crowdturfing
17:10 ~ 17:30	Challenges, Tools and Conclusion

Conceptual Level of Tutorial Theme



amazonmechanical turk
Artificial Artificial Intelligence

Your Account | HITs | Qualifications

Introduction | **Dashboard** | **Status** | **Account Settings**

Mechanical Turk is a marketplace for work.
We give businesses and developers access to an on-demand, scalable workforce.
Workers select from thousands of tasks and work whenever it's convenient.

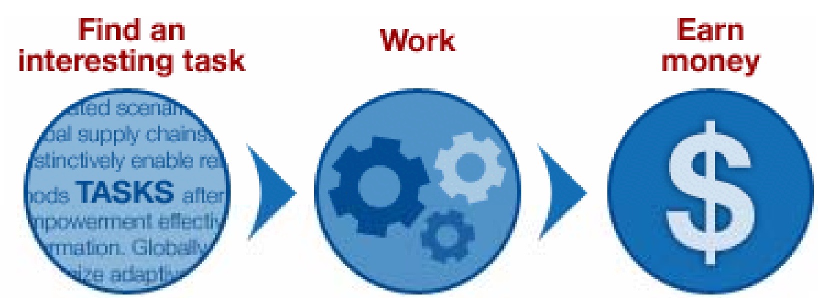
244,150 HITs available. View them now.

Make Money
by working on HITs

HITs - *Human Intelligence Tasks* - are individual tasks that you work on. Find HITs now.

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work



Find HITs Now

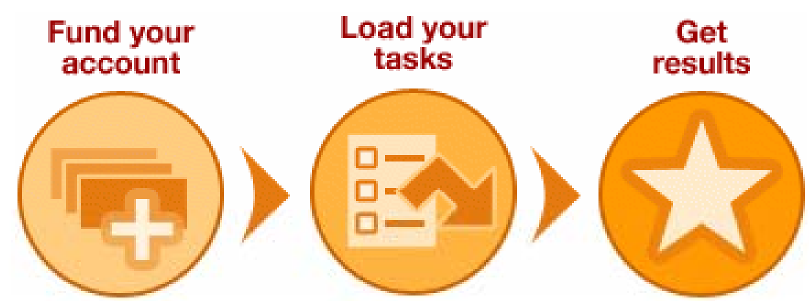
or learn more about being a **Worker**

Get Results
from Mechanical Turk Workers

Ask workers to complete HITs - *Human Intelligence Tasks* - and get results using Mechanical Turk. Register Now

As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results



Get Started

CrowdFlower Solutions Contributors Company Blog Docs Login

The World's Largest Workforce

Instantly hire millions of people to collect, filter, and enhance your data.

Business Data

Data collected at scale

The accuracy of in-house teams, the cost advantage of the crowd

Senti

Sentiment Analysis

Fast, accurate human review of user-generated social media content.

Contributors & Channels

Interested in completing microtasks or displaying a task wall to your user base?

On-Demand
Pay for only what you need when you need it.

Accurate
Guaranteed quality with rich analytics.

Fast
100x faster than traditional methods.

Experienced
Creating crowdsourcing solutions since 2007.

Crowdturfing (Crowdsourcing + Astroturfing)

- Definition of crowdturfing: masses of cheaply paid shills can be organized to spread malicious URLs in social media, form artificial grassroots campaigns ("astroturf"), and manipulate search engines.

- A Multimillion-dollar industry in Chinese crowdsourcing sites
  - 90% crowdturfing tasks [MIT Technology Review]

- 70~95% crowdturfing tasks at several U.S. crowdsourcing sites [Wang et al., WWW 2012]

Website	Campaigns	% Crowdturfing	Tasks	$ per Subm.
Amazon Turk (US)	41K	12%	2.9M	$0.092
ShortTask* (US)	30K	95%	527K	$0.096
MinuteWorkers (US)	710	70%	10K	$0.241
MyEasyTask (US)	166	83%	4K	$0.149
Microworkers (US)	267	89%	84K	$0.175

Wang et al. WWW 2012

Targeted Crowdsourcing Sites

- Eastern crowdsourcing sites
  - Zhubajie (ZBJ)
  - Sandaha (SDH)

- Western crowdsourcing sites
  - Microworkers.com
  - ShortTask.com
  - Rapidworkers.com

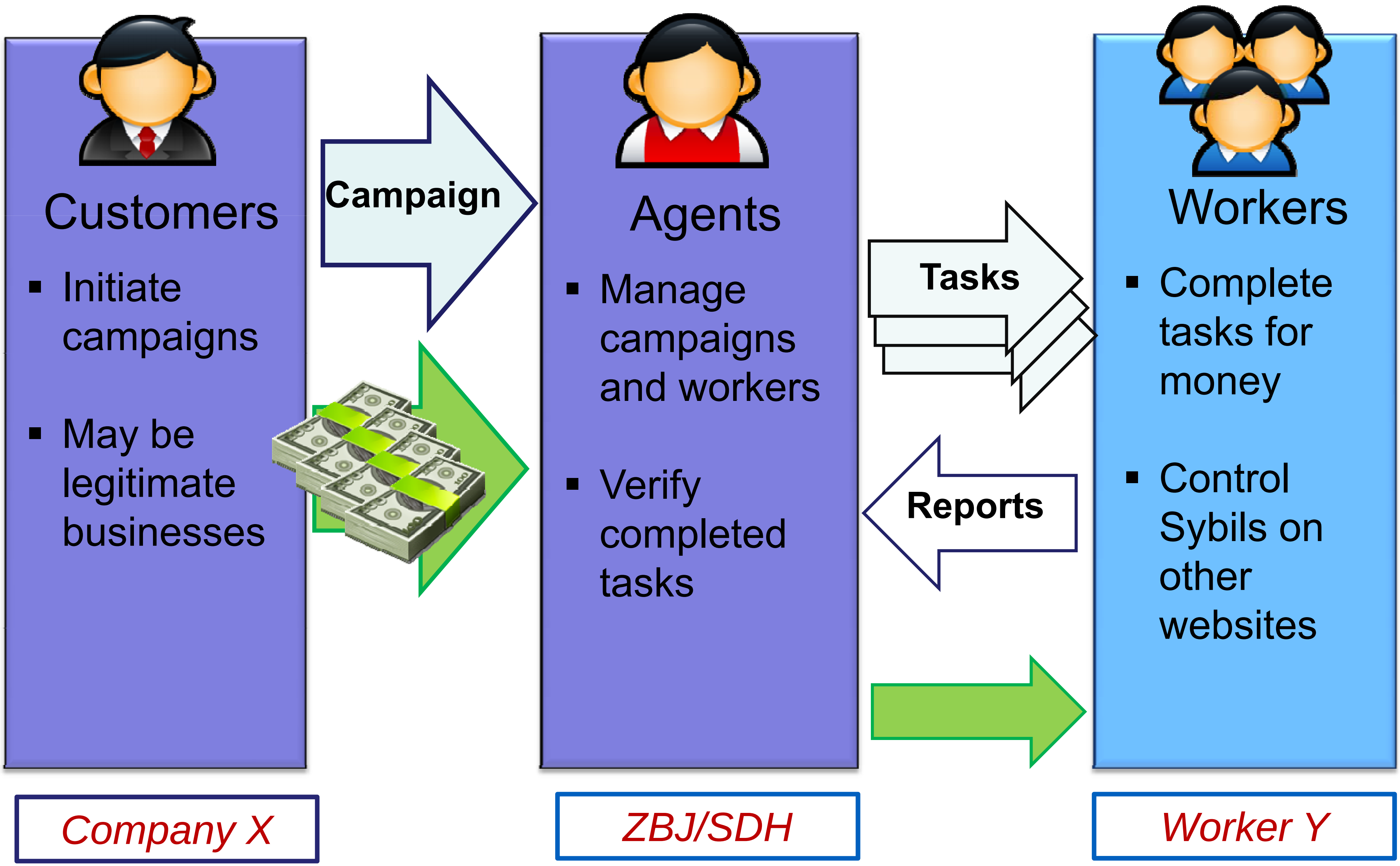
Eastern Crowdsourcing Sites

Crowdturfing Sites

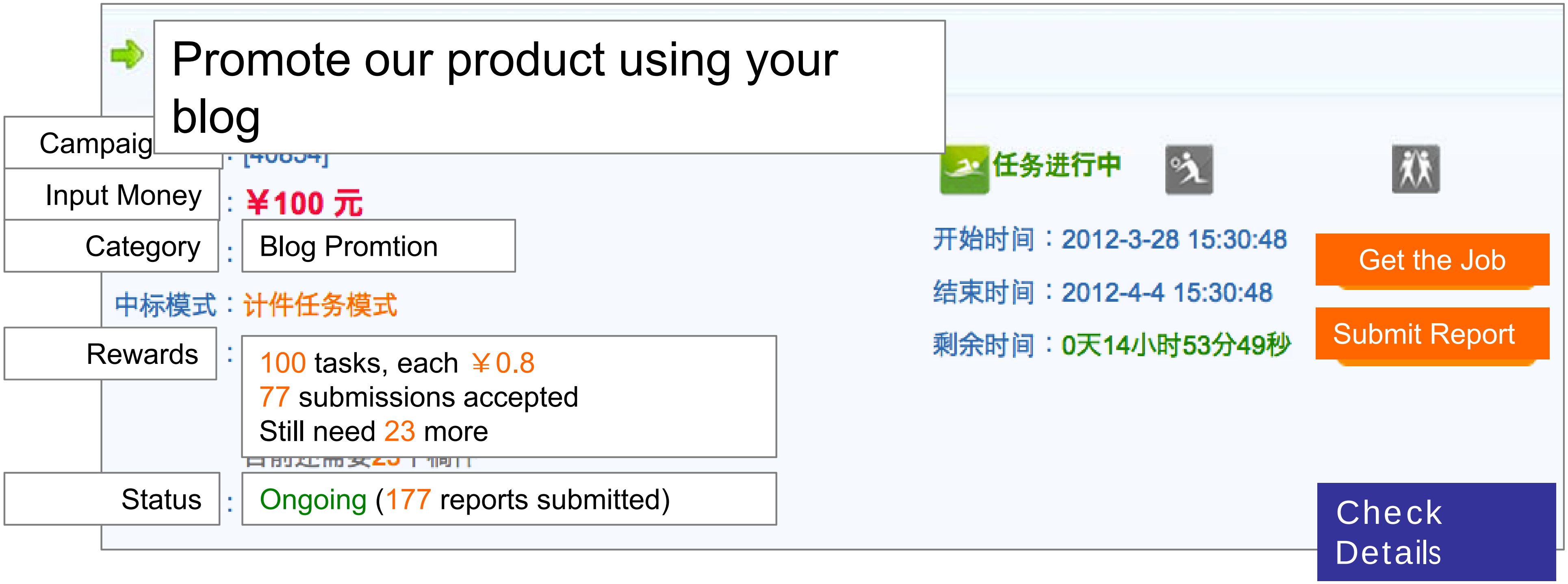
- Focus on the two largest sites
  - Zhubajie (ZBJ)
  - Sandaha (SDH)
- Crawling ZBJ and SDH
  - Details are completely open
  - Complete campaign history since going online
    - ZBJ 5-year history
    - SDH 2-year history

Wang, G., Wilson, C., Zhao, X., Zhu, Y., Mohanlal, M., Zheng, H., and Zhao, B. Y. Zhao. Serf and turf: crowdturfing for fun and profit. In WWW, 2012.

Crowdturfing Workflow



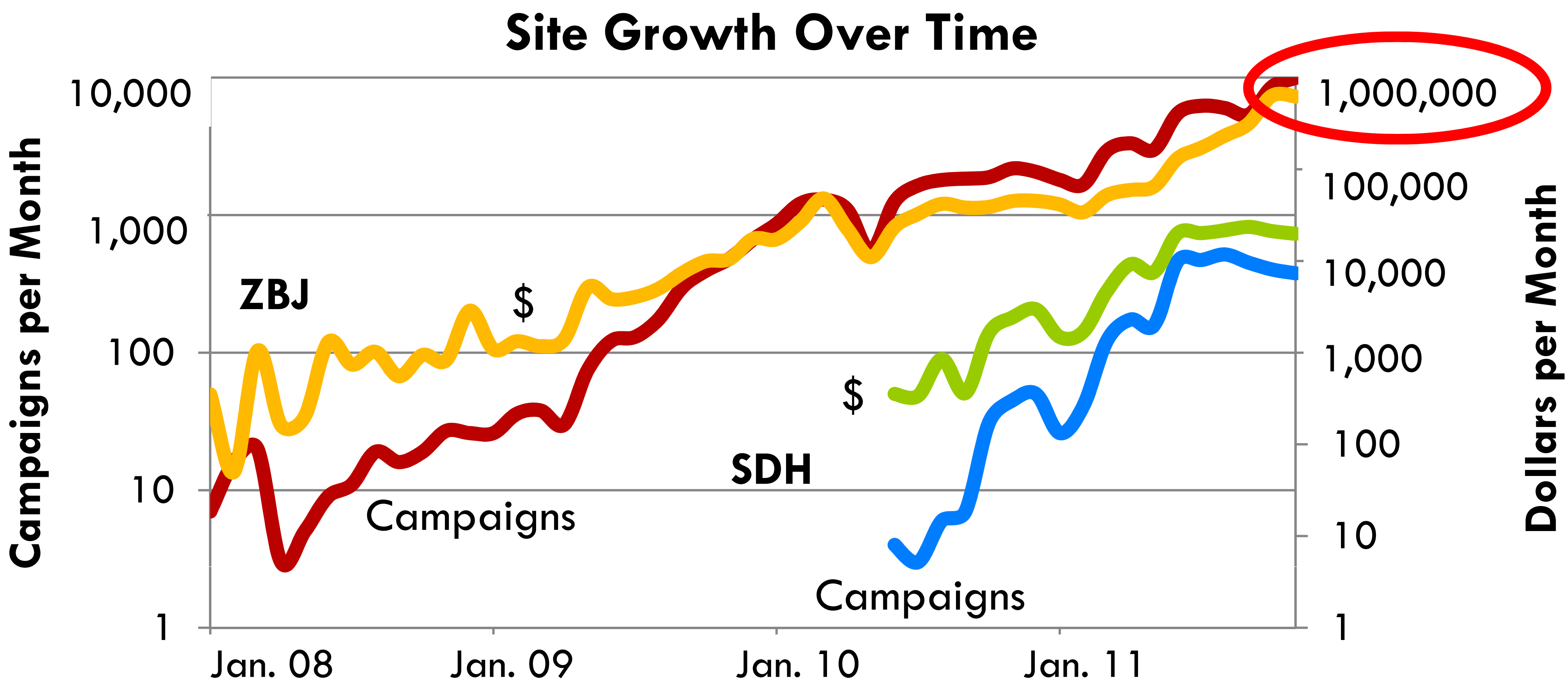
Campaign Information



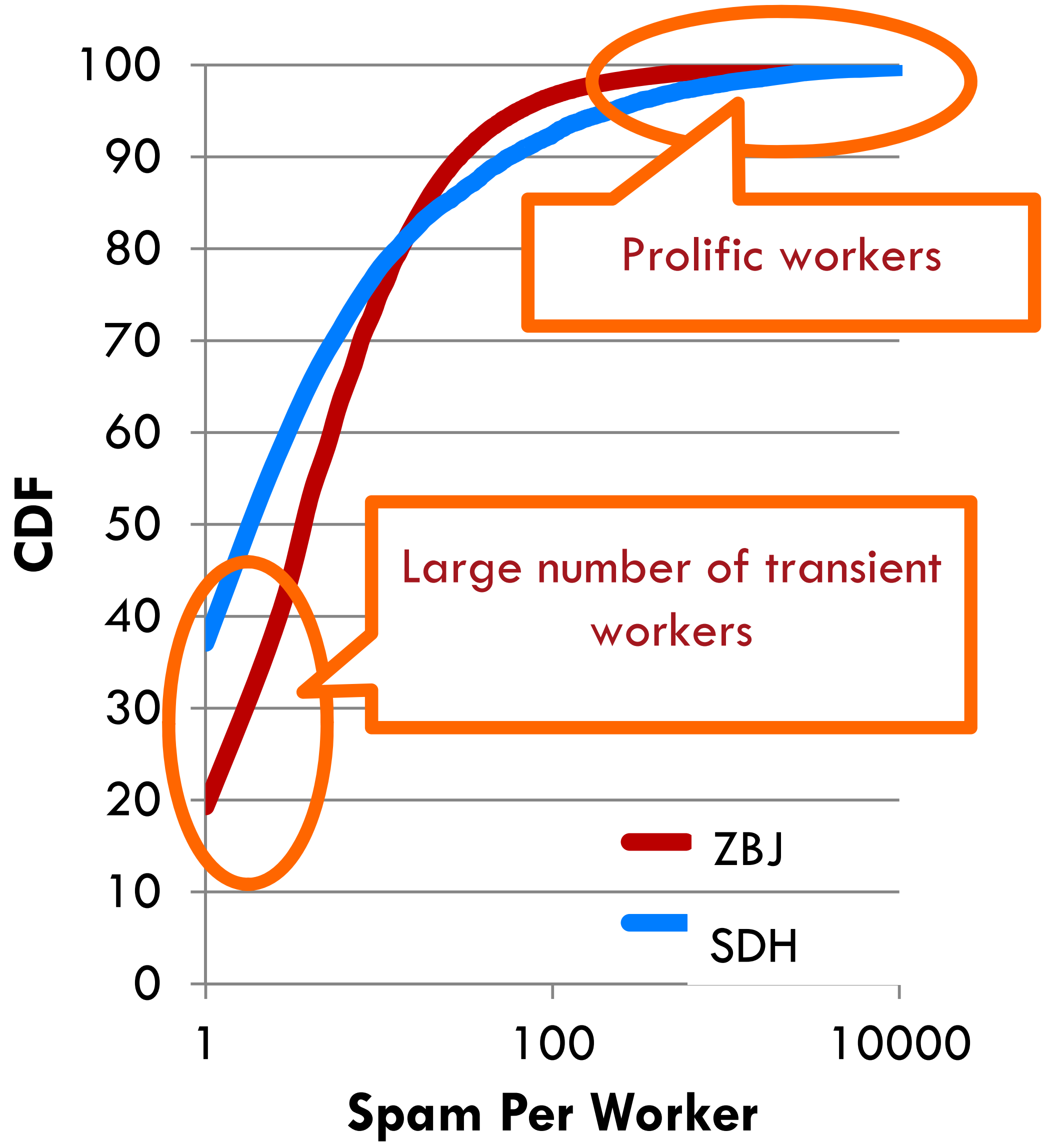
Report generated by workers

High Level Statistics

Site	Active Since	Total Campaigns	Workers	Reports	$ for Workers	$ for Site
ZBJ	Nov. 2006	76K	169K	6.3M	$2.4M	$595K

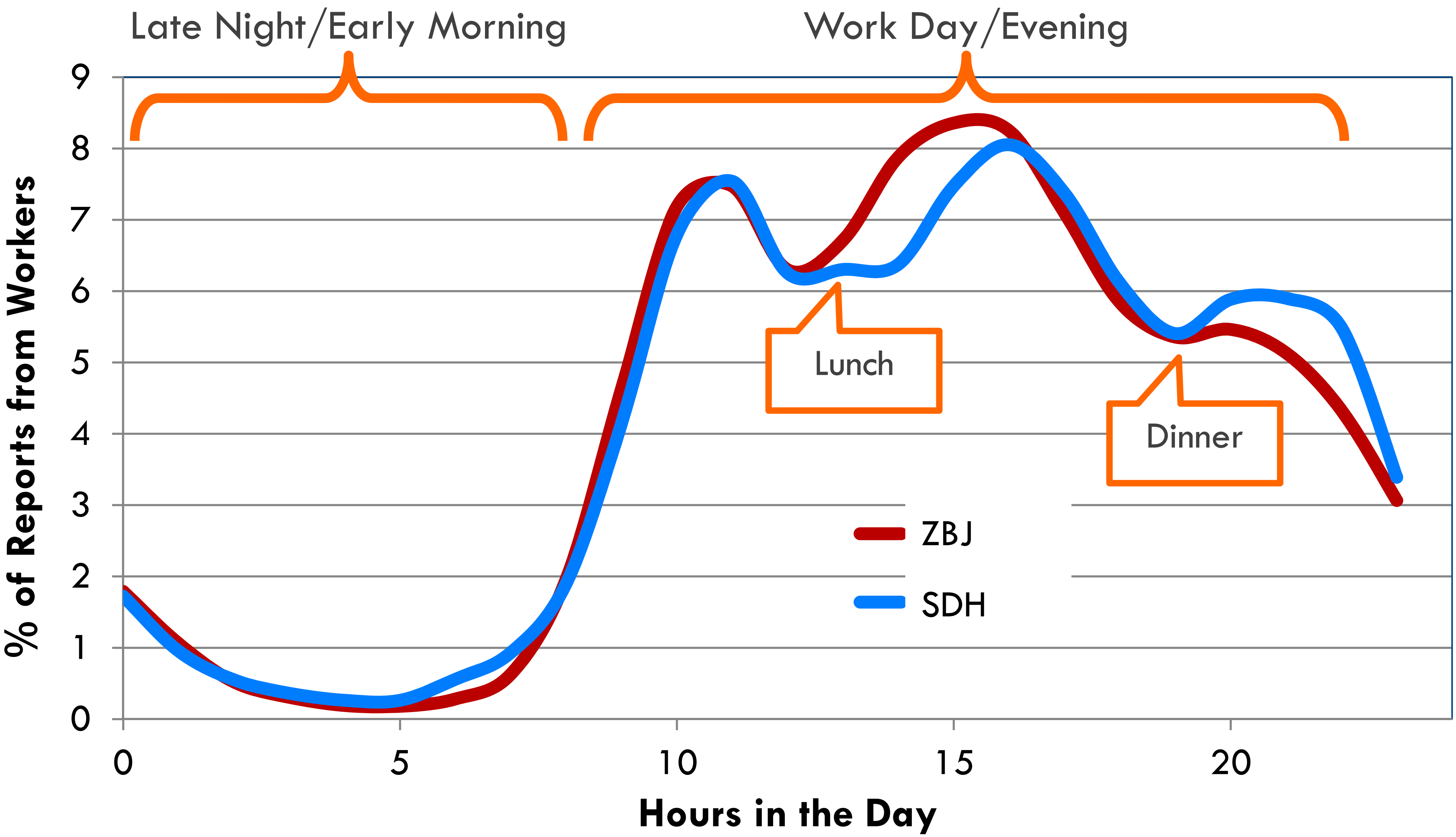


Spam Per Worker



- Transient workers
  - Makes up majority of a diverse worker population
- Prolific workers
  - Major force of spam generation

Are Workers Real People?



Campaign Types

Top 5 Campaign Types on ZBJ

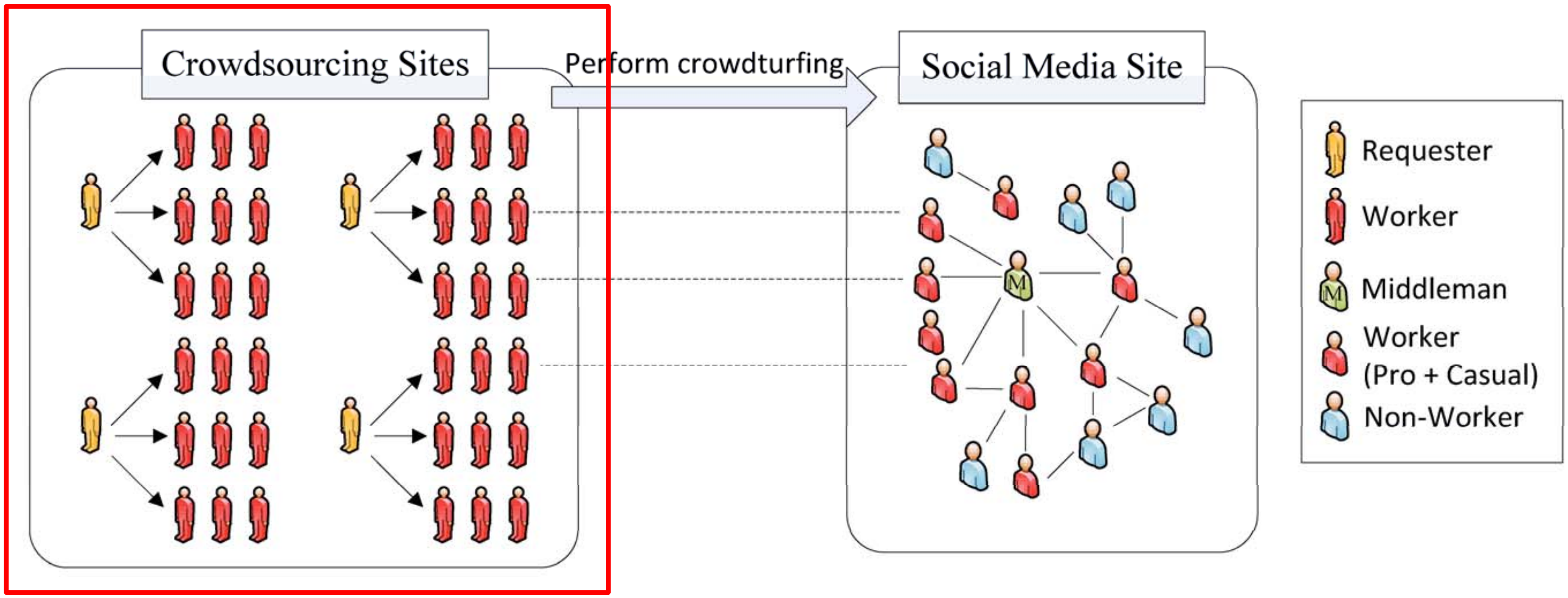
Campaign Target	# of Campaigns	$ per Campaign	$ per Spam	Monthly Growth
Account Registration	29,413	$71	$0.35	16%
Forums	17,753	$16	$0.27	19%
Instant Message Groups	12,969	$15	$0.70	17%
Microblogs (e.g. Twitter/Weibo)	4061	$12	$0.18	47%
Blogs	3067	$12	$0.23	20%

- Most campaigns are spam generation
- Highest growth category is microblogging
  - Weibo: increased by 300% (200 million users) in a single year (2011)
  - $100 → audience of 100K Weibo users

Western Crowdsourcing Sites

Research Goal and Framework

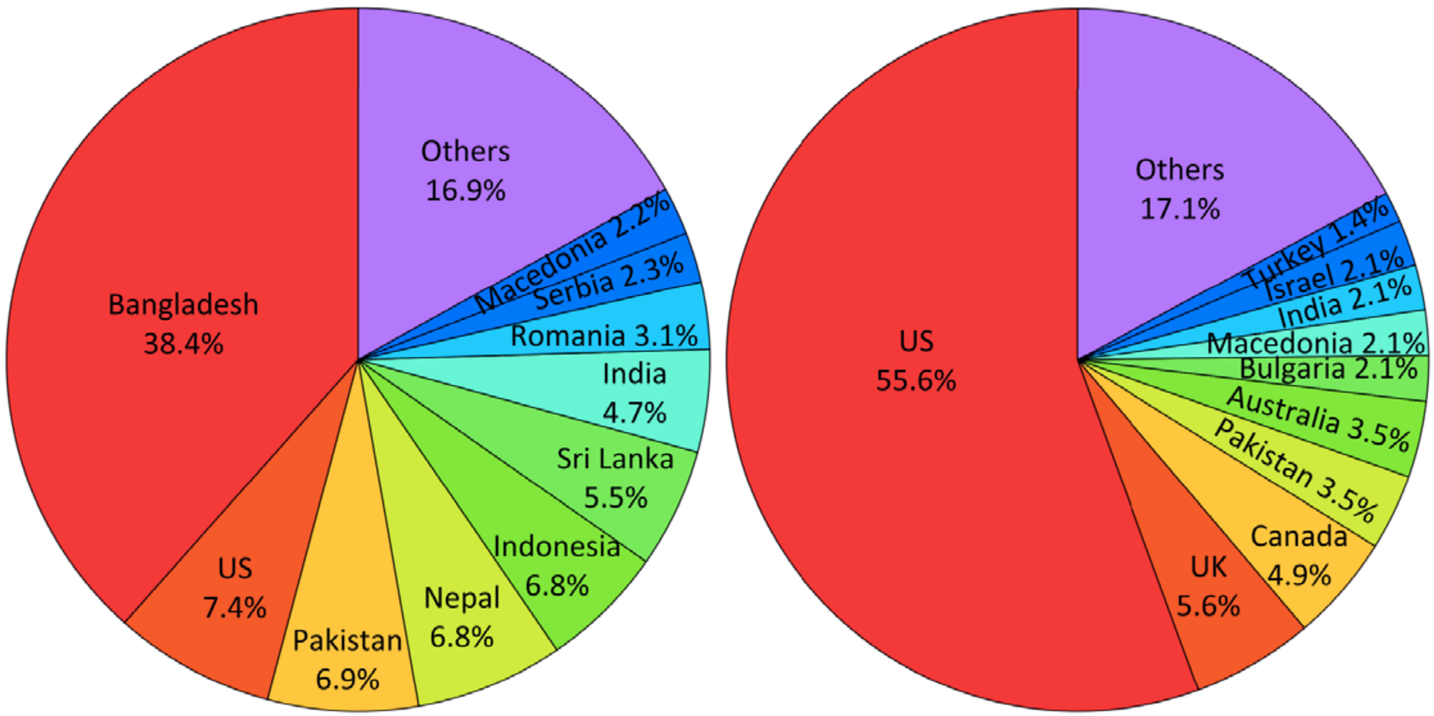
- Goal: reveal the underlying ecosystems of crowdturfers



- In crowdsourcing sites
  - Who are these participants?
  - What are their roles?
  - What types of campaigns are they engaged in?

Lee, K., Tamilarasan, P., and Caverlee, J. Crowdturfers, Campaigns, and Social Media: Tracking and Revealing Crowdsourced Manipulation of Social Media. In *ICWSM*, 2013.

Requesters and Workers



(a) Workers (b) Requesters

- Collected and analyzed 144 requesters' profiles and 4,012 workers' profiles in a Western crowdsourcing site, Microworkers.com
- Major portion of the workers are from the developing countries
- 70% of all requesters are from the English-speaking countries
  - United States, UK, Canada, and Australia.
- Surprisingly, the workers have done about 3 million tasks and have earned a half million dollars

Analysis of Crowdturfing Tasks

- Dataset: sampled 505 tasks containing 63,042 jobs from three Western crowdsourcing sites such as Microworkers.com, ShortTask.com and Rapidworkers.com.
- Five groups of the Tasks
  - Social Media Manipulation [56%]:
    - Workers to target social media
  - Sign Up [26%]:
    - Workers to sign up on a website for several reasons (e.g., to increase the user pool, and promote advertisements)
  - Search Engine Spamming [7%]:
    - Workers to search for a certain keyword on a search engine, and then click the specified link
  - Vote Stuffing [4%]:
    - Workers to cast votes
  - Miscellany [7%]:
    - Some other activity

Vote Stuffing

Music Awards: Sign up + Vote for Tommy

1. Go to www.vcmusicawards.com
2. Register to vote
3. Go to the BEST BLUES BAND catagory
4. Vote for TOMMY MARSH and BAD DOG

Top Rated

Tommy Marsh & Bad Dog

320 votes

D.on Darox & The Melody Joy Bakers

104 votes

50 Sticks of Dynamite

22 votes

R&B Bombers

19 votes

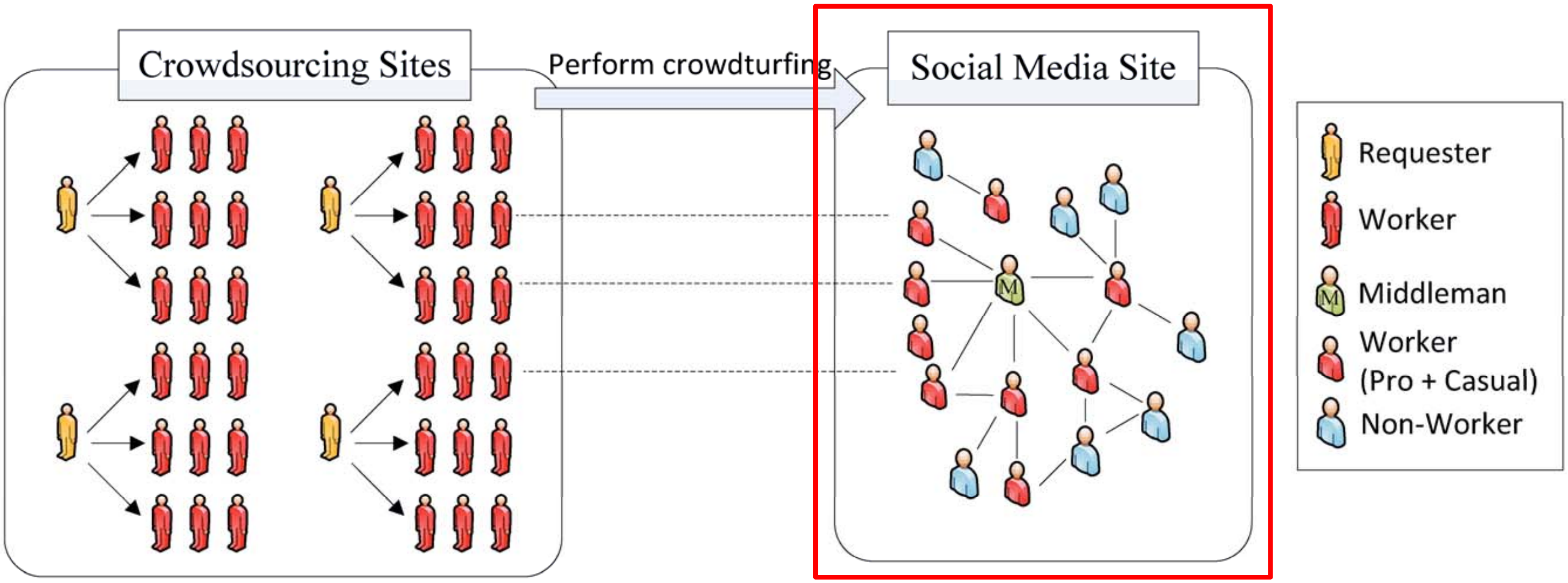
The Front Street Prophets

7 votes

Tommy Marsh & Bad Dog

Best Blues Band Nominee

Research Questions in Social Media



- By linking crowdturfing tasks and participants on crowdsourcing sites to social media
  - Can we uncover the implicit power structure of crowdturfers?
  - Can we automatically distinguish between the behaviors of crowdturfers and regular social media users?

Linking Crowdsourcing Workers to Social Media

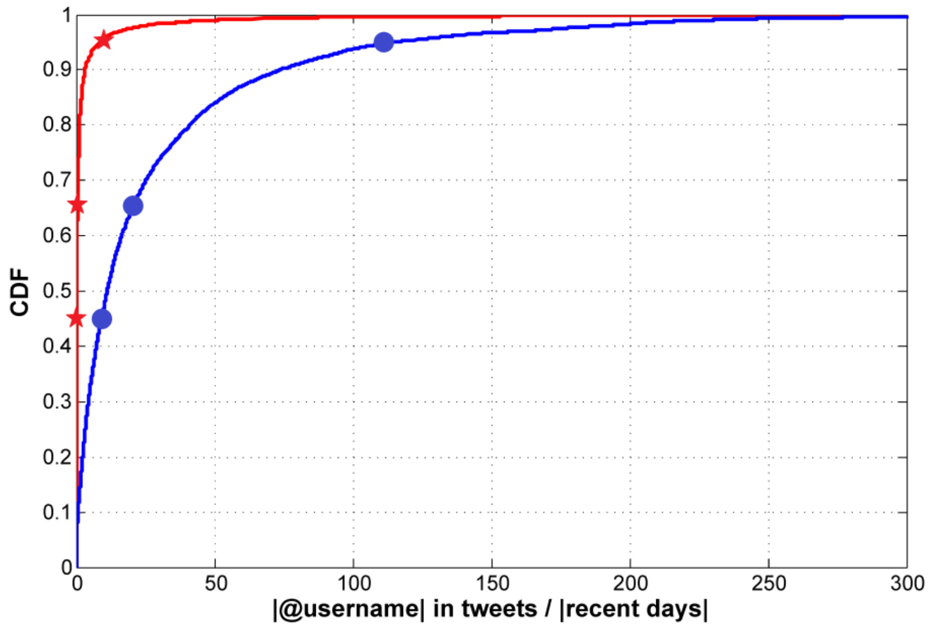
- 65 out of 505 tasks (campaigns) targeted Twitter.
  - Tweeting about a link
  - Following a twitter user

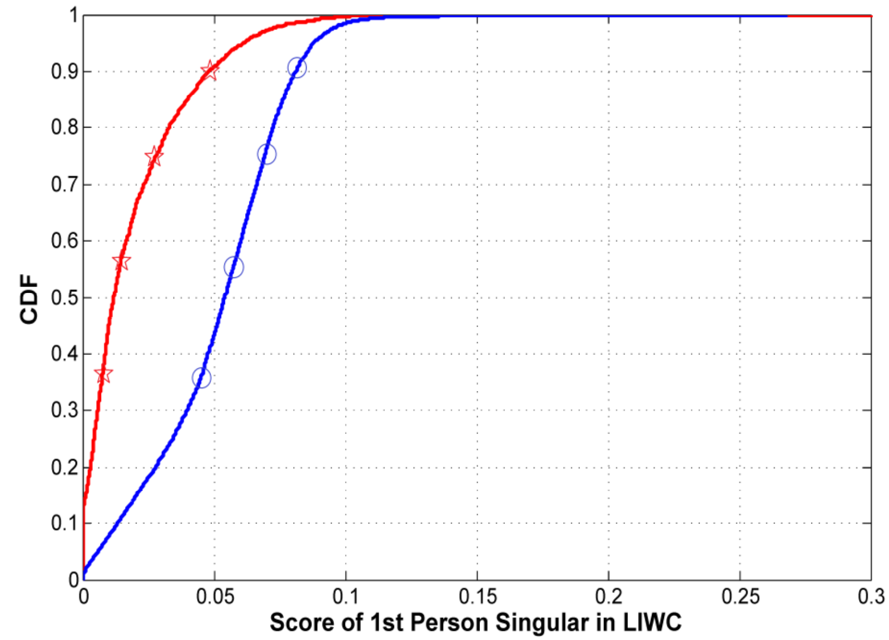
- Twitter Dataset

Class	$\lvert$User Profiles$\rvert$	$\lvert$Tweets$\rvert$
Workers	2,864	364,581
Non-Workers	9,878	1,878,434

Analysis of Twitter Workers

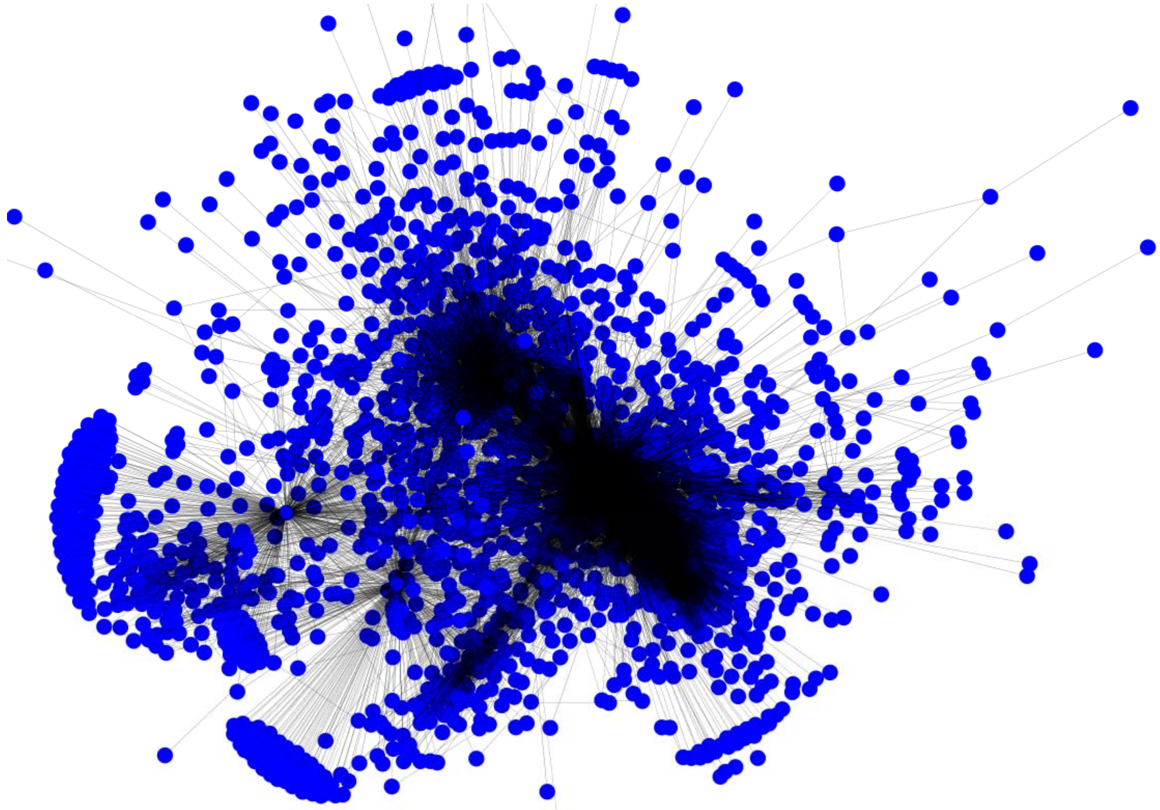
- Activity and linguistic characteristics (by LIWC)





- workers rarely communicate with other users via @username
- workers are less personal in the messages they post than non-workers

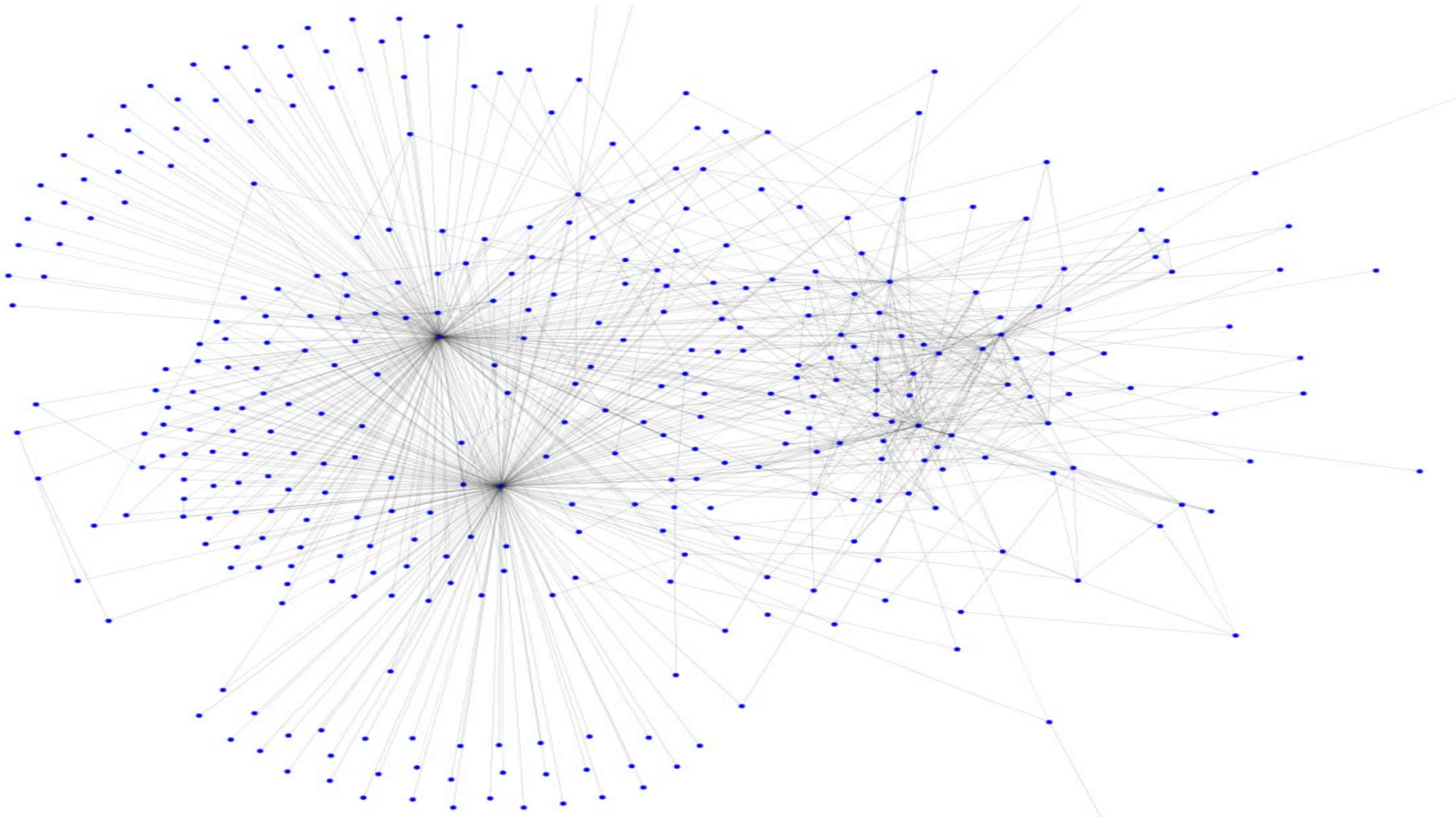
Network Structure of Twitter Workers



- Twitter workers on average are densely connected to each other.
- The graph density of the workers is higher than the average graph density of Twitter users.

Professional Workers

- Definition: participated in three or more tasks targeting Twitter.
- Surprisingly, graph density of 187 professional workers is even higher than all workers' graph density



Middlemen

- Definition of Middlemen: Whose messages were often retweeted by the professional workers. These middlemen are the message creators.

- Top-10 Middlemen

Middleman	|Pro-Workers|	|Followings|	|Followers|
0boy	139	847	108,929
louiebaur	95	285	68,772
hasai	63	6,360	41,587
soshable	57	956	22,676
virtualmember	56	5,618	5,625
scarlettmadi	55	5,344	26,439
SocialPros	54	10,775	22,985
cqlivingston	54	6,377	28,556
huntergreene	49	27,390	25,207
TKCarsitesInc	48	1,015	18,661

•Most of the middlemen are interested in social media strategy, social marketing and SEO.

•Several middlemen opened their location as Orange County, CA.

•Some of them also often retweeted other middlemen's messages.

Detecting Crowd Workers

- Twitter Dataset:

Class	|User Profiles|	|Tweets|
Workers	2,864	364,581
Non-Workers	9,878	1,878,434

- Feature Categories
  - User Demographics: account age, and other descriptive information about the user
  - User Friendship Networks: number of followers, following and bi-directional friends, etc
  - User Activity: number of posted tweets, number of links in tweets, etc
  - User Content : personality features (LIWC), content similarity, etc

- Top-10 Features (by chi-square)

Feature	Workers	Non-workers
|links| in tweets / |tweets|	0.696	0.142
|tweets| / |recent days|	4	37
|@username| in tweets / |recent days|	2	28
the number of posted tweets per day	3	21
|rt| in tweets / |tweets|	0.7	9.7
Swearing in LIWC	0.001	0.009
|links| in RT tweets / |RT tweets|	0.589	0.142
Anger in LIWC	0.003	0.012
Total Pronouns in LIWC	0.054	0.107
1st Person Singular in LIWC	0.019	0.051

Detecting Crowd Workers (Cont'd)

- Performance Results (by 10-fold cross-validation)

Classifier	Accuracy	F1	AUC	FNR	FPR
Random Forest	93.26%	0.966	0.955	0.036	0.174

- Consistency of Worker Detection over Time (a month later)

Class	User Profiles	Tweets
Workers	368	40,344

Classifier	Accuracy	F1	FNR
Random Forest	94.3%	0.971	0.057

This positive experimental result shows that their classification approach is promising to find new workers in the future

So far…Crowdturfing

- Eastern crowdsourcing sites
  - Zhubajie (ZBJ)
  - Sandaha (SDH)

- Western crowdsourcing sites
  - Microworkers.com
  - ShortTask.com
  - Rapidworkers.com

Reference List

- Motoyama, M., McCoy, D., Levchenko, K., Savage, S., and Voelker, G. M. Dirty jobs: the role of freelance labor in web service abuse. In *Proceedings of the 20th USENIX conference on Security (SEC)*, 2011.
- Lee, K., Tamilarasan, P., and Caverlee, J. Crowdturfers, Campaigns, and Social Media: Tracking and Revealing Crowdsourced Manipulation of Social Media. In *ICWSM*, 2013.
- Wang, G., Wilson, C., Zhao, X., Zhu, Y., Mohanlal, M., Zheng, H., and Zhao, B. Y. Zhao. Serf and turf: crowdturfing for fun and profit. In *WWW*, 2012.
- K. Lee, S. Webb, and H. Ge. The Dark Side of Micro-Task Marketplaces: Characterizing Fiverr and Automatically Detecting Crowdturfing. In *ICWSM*, 2014.

Schedule

14:00 ~ 14:10	Introduction to Social Media Threats (Social Spam, Campaigns, Misinformation and Crowdturfing)
14:10 ~ 14:55	Social Spam
14:55 ~ 15:30	Campaigns
15:30 ~ 16:00	Break
16:00 ~ 16:30	Misinformation
16:30 ~ 17:10	Crowdturfing
17:10 ~ 17:30	Challenges, Tools and Conclusion

Open Research Challenges

- Need for large, accurate, up-to-date data sets
  - APIs
  - Hard crawling
  - Shared datasets
  - Purchasing data (e.g., Gnip)
  - Data grant or know an insider
- Labeling
  - Manual labeling
  - Use crowd wisdom
  - Get labeled data from a social media site
  - Blacklist

Open Research Challenges

- Integration of multiple techniques for data processing and modeling
  - Big data analysis, machine learning (data mining), information retrieval, visualization, etc
- Interdisciplinary research for analysis
  - computer science, social science, psychology, etc
- Arms race (endless battle)
  - Spammers and malicious users change their behaviors or use new techniques to avoid existing detection approaches
  - Spammers and malicious users move to another site

Useful Tools

- Machine learning
  - Weka: http://www.cs.waikato.ac.nz/ml/weka/
  - scikit-learn: http://scikit-learn.org/stable/
  - LingPipe (linguistic analysis): http://alias-i.com/lingpipe/

- Visualization
  - Matplotlib: http://matplotlib.org/
  - Gephi: https://gephi.org/
  - Graphviz: http://www.graphviz.org/

Useful Tools

- Big data analysis and visualization
  - Hadoop (MapReduce): http://hadoop.apache.org/
  - Pig: https://pig.apache.org/
  - Hive: https://hive.apache.org/
  - Cascalog: http://cascalog.org/
  - Giraph: https://giraph.apache.org/
- Scalable machine learning
  - Mahout: https://mahout.apache.org/
- Large scale stream processing
  - Storm: http://storm.incubator.apache.org/
  - Summingbird: https://github.com/twitter/summingbird

Conclusion

- We covered four social media threats
  - Social Spam
  - Campaigns
  - Misinformation
  - Crowdturfing
- We focused on countermeasures and their experimental results
- Tutorial slides:
  - http://digital.cs.usu.edu/~kyumin/tutorial/www2014.html

All Reference List

- Grier, C., Thomas, K., Paxson, V., and Zhang, M. @spam: the underground on 140 characters or less. In CCS, 2010.
- Lee, S., and Kim, J. WarningBird: Detecting suspicious URLs in Twitter stream. In *NDSS*, 2012.
- Ghosh, S., Viswanath, B., Kooti, F., Sharma, N. K., Korlam, G., Benevenuto, F., Ganguly, N., and Gummadi, P. K. Understanding and combating link farming in the twitter social network. In *WWW*, 2012.
- Benevenuto, F., Rodrigues T., Almeida V., Almeida, J., and Gonçalves, M. Detecting spammers and content promoters in online video social networks. In *SIGIR*, 2009.
- Lee, K., Eoff, B., and Caverlee, J. Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter. In *ICWSM*, 2011.
- Aggarwal, A., Almeida, J., and Kumaraguru, P. Detection of spam tipping behaviour on foursquare. In *WWW Companion*, 2013.
- Lee., K., Caverlee., J., and Webb, S. Uncovering Social Spammers: Social Honeypots + Machine Learning. In *SIGIR*, 2010.
- Wang, G., Mohanlal, M., Wilson, C., Wang, X., Metzger, M. J., Zheng, H., and Zhao, B. Y. Social Turing Tests: Crowdsourcing Sybil Detection. In *NDSS*, 2013.
- Tan, E., Guo, L., Chen, S., Zhang, X., and Zhao, Y. UNIK: Unsupervised Social Network Spam Detection. In *CIKM*, 2013
- Lee, K., Kamath, K., and Caverlee, J. Combating Threats to Collective Attention in Social Media: An Evaluation. In *ICWSM*, 2013.
- Gao, H., Hu J., Wilson, C., Li, Z., Chen, Y., and Zhao, B. Detecting and characterizing social spam campaigns. In *IMC*, 2010.
- Lee, K., Caverlee, J., Cheng, Z., and Sui, D. Content-Driven Detection of Campaigns in Social Media. In *CIKM*, 2011

All Reference List

- Lee, K., Caverlee, J., Cheng, Z., and Sui, D. Campaign Extraction from Social Media. In *ACM TIST, Vol. 5, No. 1*, January 2014.
- Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Flammini, A., and Menczer, F. Detecting and Tracking Political Abuse in Social Media. In *ICWSM*, 2011.
- Mukherjee, A., Liu, B., and Glance, N. Spotting fake reviewer groups in consumer reviews. In *WWW*, 2012.
- Castillo, C., Mendoza, M., and Poblete, B. Information credibility on twitter. In *WWW*, 2011.
- Yang, F., Liu, Y., Yu, X., and Yang, M. Automatic detection of rumor on Sina Weibo. In *SIGKDD Workshop on Mining Data Semantics*, 2012.
- Gupta, A., Lamba, H., Kumaraguru, P., and Joshi, A. Faking Sandy: characterizing and identifying fake images on Twitter during Hurricane Sandy. In *WWW Companion*, 2013.
- Xia, X., Yang, X., Wu, C., Li, S., and Bao, L. Information credibility on twitter in emergency situation. In *Proceedings of the 2012 Pacific Asia conference on Intelligence and Security Informatics (PAISI)*, 2012.
- Motoyama, M., McCoy, D., Levchenko, K., Savage, S., and Voelker, G. M. Dirty jobs: the role of freelance labor in web service abuse. In *Proceedings of the 20th USENIX conference on Security (SEC)*, 2011.
- Lee, K., Tamilarasan, P., and Caverlee, J. Crowdturfers, Campaigns, and Social Media: Tracking and Revealing Crowdsourced Manipulation of Social Media. In *ICWSM*, 2013.
- Wang, G., Wilson, C., Zhao, X., Zhu, Y., Mohanlal, M., Zheng, H., and Zhao, B. Y. Zhao. Serf and turf: crowdturfing for fun and profit. In *WWW*, 2012.
- K. Lee, S. Webb, and H. Ge. The Dark Side of Micro-Task Marketplaces: Characterizing Fiverr and Automatically Detecting Crowdturfing. In *ICWSM*, 2014.

Thanks to…

- All authors in the reference list for sharing their presentation slides.

Thank you