Social Spam, Campaigns, Misinformation and Crowdturfing

Kyumin Lee  James Caverlee  Calton Pu
Utah State Univ.  Texas A&M Univ.  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

## Location-based Services
- foursquare
- Yelp
- Google Latitude
- Gowalla

## Crowd-based services
- CrowdFlower
- Kickstarter
- IndieGoGo
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
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

- ...

Large-Scale Social Systems: Challenges and Research Approach
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
100.143.232.12

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 woman.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-Website x
justinjkt11558@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 upsets 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

Results for #DrakeCriesWhen

CherryKornblut
No way. She pulls this again!! bit.ly/oFalmo #gritosmexicanos #DrakeCriesWhen #FastFoodAddiction Glen Rice
18 seconds ago

ApyrlHemmelgarn
I wonder if this really works bit.ly/oFalmo Glen Rice
#UKnowHUhungryWhen #gritosmexicanos #DrakeCriesWhen #FastFoodAddiction
18 seconds ago

UnSchwegel6358
Anybody know is this really works!!?? bit.ly/oFalmo #gritosmexicanos Glen Rice #UKnowHUhungryWhen #DrakeCriesWhen #FastFoodAddiction
18 seconds ago

DebraRozzi7831
No way. She pulls this again!! bit.ly/oFalmo #gritosmexicanos #DrakeCriesWhen #FastFoodAddiction Glen Rice
17 seconds ago

MaryaLiccketto9
Omg...Is this real?? bit.ly/oFalmo #gritosmexicanos #DrakeCriesWhen #FastFoodAddiction Glen Rice
16 seconds ago
Campaigns

Astroturfing

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

Fake review campaign

1 of 1 people found the following review helpful:

5/5 Stars 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...

3 of 8 people found the following review helpful:

5/5 Stars 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, ...

5 of 5 people found the following review helpful:

5/5 Stars 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 ...

Political campaign

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

How true? This network graph shows the connections between 6,276 accounts that used the hashtag #pimp 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?"

Wang et al. WWW 2012

<table>
<thead>
<tr>
<th>Website</th>
<th>Campaigns</th>
<th>% Crowdfunding</th>
<th>Tasks</th>
<th>$ per Subm.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon Turk (US)</td>
<td>41K</td>
<td>12%</td>
<td>2.9M</td>
<td>$0.092</td>
</tr>
<tr>
<td>ShortTask* (US)</td>
<td>30K</td>
<td>95%</td>
<td>527K</td>
<td>$0.096</td>
</tr>
<tr>
<td>MinuteWorkers (US)</td>
<td>710</td>
<td>70%</td>
<td>10K</td>
<td>$0.241</td>
</tr>
<tr>
<td>MyEasyTask (US)</td>
<td>166</td>
<td>83%</td>
<td>4K</td>
<td>$0.149</td>
</tr>
<tr>
<td>Microworkers (US)</td>
<td>267</td>
<td>89%</td>
<td>84K</td>
<td>$0.175</td>
</tr>
</tbody>
</table>
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

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 (SMSC) program. It's an attempt to get better at both detecting and
Misinformation (Fake)

DC Maryland Virginia
@DMVFollowera

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

Fake Images

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

I TOLD Y’ALL! Shark on the highway in New Jersey @maxthewanted would appreciate this. #Hurricane pic.twitter.com/kaYMjWzT
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]
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

“Dairy giant Mengniu in smear scandal”

- 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

Warning: Company Y’s baby formula contains dangerous hormones!
Conceptual Level of Tutorial Theme

[1] Social Spam (Individual Spammer/Content Polluter)

[2] Campaigns (Groups of users)


1. Get Info
2. Feed Misinformation
3. Attention

Popular keywords or topics
Introduction to Social Media Threats (Social Spam, Campaigns, Misinformation and Crowdturfing)

Social Spam

Campaigns

Break

Misinformation

Crowdturfing

Challenges, Opportunities and Tools in Social Spam, Campaigns, Misinformation and Crowdturfing Research
Conceptual Level of Tutorial Theme

[1] Social Spam (Individual Spammer/Content Polluter)

[2] Campaigns (Groups of users)


1. Get Info
2. Feed Misinformation
3. Attention

Popular keywords or topics

Social Spam (Individual Spammer/Content Polluter)
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**
- **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**
- …
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
  - 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

![Graphs showing number of tweets over days for URIBL and Google malware, with red indicating lag and blue indicating lead.](image)
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.

blackraybansunglasses.com

- 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

July 11, 2011
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 $\Rightarrow$ malicious, active $\Rightarrow$ 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

Example of Video Spam

Polska-Czechy 2:1 wszytskie bramki

Advertises Pornography

Cartoon
Example of Promotion

Eric and the Army of the Phoenix (1/5)

An incredible but true story: Spanish authorities prosecute child for terrorism when he e-mails companies requesting labelling in Catalan language, using Phoenix monicker from Harry Potter books.

Poli (more)
Step 3. 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: $\chi^2$ ranking**

<table>
<thead>
<tr>
<th>Attribute Set</th>
<th>Top 10</th>
<th>Top 20</th>
<th>Top 30</th>
<th>Top 40</th>
<th>Top 50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video</td>
<td>9</td>
<td>18</td>
<td>25</td>
<td>30</td>
<td>36</td>
</tr>
<tr>
<td>User</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>SN</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>
Distinguishing classes of users

Promoters target unpopular content

Spammers target popular content
Step 4. Classification Approach

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

Flat

Hierarchical

Promoters
Spammers
Legitimates

Promoters
Non-promoters

Light
Heavy
Spammers
Legitimates
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

<table>
<thead>
<tr>
<th>True</th>
<th>Promoter</th>
<th>Spammer</th>
<th>Legitimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Promoter</td>
<td>96.13%</td>
<td>3.87%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Spammer</td>
<td>1.40%</td>
<td>56.69%</td>
<td>41.91%</td>
</tr>
<tr>
<td>Legitimate</td>
<td>0.31%</td>
<td>5.02%</td>
<td>94.66%</td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Promoter</th>
<th>Non-Promoter</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>True</strong></td>
<td>Promoter</td>
<td>Non-Promoter</td>
</tr>
<tr>
<td>Promoter</td>
<td>92.26%</td>
<td>7.74%</td>
</tr>
<tr>
<td>Non-Promoter</td>
<td>0.55%</td>
<td>99.45%</td>
</tr>
</tbody>
</table>
Distinguishing Spammers from Legitimate users

- **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
Foursquare Spam Tips

- Tips unrelated to Venue
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

<table>
<thead>
<tr>
<th>Category</th>
<th>( \chi^2 ) rank</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>User Attributes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Number of Tips</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Ratio of Check-ins and Tips</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Number of Check-ins</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Number of Badges</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Number of Mayorships</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Ratio of Check-ins and Badges</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Number of Photos posted</td>
<td></td>
</tr>
<tr>
<td><strong>Social Attributes</strong></td>
<td>6</td>
<td>Number of Friends</td>
</tr>
<tr>
<td><strong>Content Attributes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Similarity score of Tips</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Number of URLs posted</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Average number of words in Tips</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Average number of characters in Tips</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Ratio of number of likes and number of Tips</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Average number of spam words in Tips</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Average number of phone-numbers posted in Tips</td>
<td></td>
</tr>
</tbody>
</table>
## Classification Results

<table>
<thead>
<tr>
<th>Classification Algorithm</th>
<th>Precision (Spam)</th>
<th>Precision (Safe)</th>
<th>Recall (Spam)</th>
<th>Recall (Safe)</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN</td>
<td>83.2%</td>
<td>86.6%</td>
<td>86.3%</td>
<td>83.5%</td>
<td>84.89%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>88.1%</td>
<td>89.2%</td>
<td>88.3%</td>
<td>85.8%</td>
<td>89.53%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>89.3%</td>
<td>90.2%</td>
<td>88.3%</td>
<td>90.3%</td>
<td>89.76%</td>
</tr>
</tbody>
</table>
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 Polluters Tempted

• Content Polluter Examples

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

Do spammers engage in link farming?

Spam-targets: Users followed by spammers

27% (15/54) of entire Twitter targeted!
Do spammers engage in link farming?

Spam-followers: Users following spammers

- spam-targets: 15M
- spam-followers: 1.38M

27% (15/54) of entire Twitter targeted!
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

82% of spam-followers are also targeted users
Are link farmers real users or spammers?

• To find out if they are spammers or real users, the researchers
  – 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

40% of spammers appear in top 20% according to Pagerank

Most of the spammers get pushed to last 10% positions based on Collusionrank
Effect on link farmers

87% of link farmers in top 2% users according to Pagerank

98% of the link farmers get pushed to last 10% positions based on Collusionrank
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

Real or fake?

Why?

Navigation Buttons

Classifying Profiles

Browsing Profiles

Screenshot of Profile (Links Cannot be Clicked)
## Experiment Overview

<table>
<thead>
<tr>
<th>Dataset</th>
<th># of Profiles</th>
<th>Test Group</th>
<th># of Testers</th>
<th>Profile per Tester</th>
</tr>
</thead>
<tbody>
<tr>
<td>Renren</td>
<td>100 100</td>
<td>Chinese Expert</td>
<td>24</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chinese Turker</td>
<td>418</td>
<td>10</td>
</tr>
<tr>
<td>Facebook US</td>
<td>32 50</td>
<td>US Expert</td>
<td>40</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US Turker</td>
<td>299</td>
<td>12</td>
</tr>
<tr>
<td>Facebook India</td>
<td>50 49</td>
<td>India Expert</td>
<td>20</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>India Turker</td>
<td>342</td>
<td>12</td>
</tr>
</tbody>
</table>
Individual Tester Accuracy

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

>80% accuracy!

Much Lower Accuracy
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

Removing inaccurate turkers can effectively reduce false negatives!

Dramatic Improvement
System Architecture

Crowdsourcing Layer

OSN Employees

Turker Selection

Very Accurate Turkers

Accurate Turkers

All Turkers

Rejected!

Social Network

User Reports

Suspicious Profiles

Flag Suspicious Users

• Continuous Quality Control
• Locate Malicious Workers
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, K., Eoff, B., and Caverlee, J. Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter. In ICWSM, 2011.
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

1. Get Info
2. Feed Misinformation
3. Attention

Popular keywords or topics

[2] Campaigns (Groups of users)

[1] Social Spam (Individual Spammer/Content Polluter)

Origin: Crowdturfing

Misinformation

Campaigns

Social Spam
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.

Build Post Similarity Graph

- A node: an individual wall post
- An edge: connect two “similar” wall posts

Go to evil.com!

Check out funny.com
Wall Post Similarity Metric

Spam wall post model:

A textual description:

A destination URL:

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

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

Establish an edge!
Extract Wall Post Campaigns

• Intuition:

\[ A \rightarrow B \rightarrow C \]

• Reduce the problem of identifying potential campaigns to identifying connected subgraphs.
Locate Spam Campaigns

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

Wall post campaign

Distributed?

YES

Malicious

NO

Benign

Bursty?

YES

NO

Benign

Benign
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
  - 3rd party tools
  - Redirection analysis
  - Keyword matching
  - URL grouping
  - 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

- Obfuscated URL: 6.3%
- Blacklisted URL: 28.0%
- Redirection Analysis: 27.9%
- Keyword matching: 1.2%
- URL grouping: 32.5%
- Manual confirmation: 0.1%

True Positives (ALL): 96.1%
False Positives: 3.9%

The validation result.
Spam Campaign Goal Analysis

- Categorize the attacks by attackers’ goals.

Bar chart showing:
- Phishing #1: for money
- Phishing #2: for info

# of Wall Posts

Categories: phishing, malware, narcotics, pharma, luxury, other
Content-driven campaign detection
### Message Level Campaign Detection

#### Messages

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

---

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”}


**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

<table>
<thead>
<tr>
<th>Approach</th>
<th>F₁</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unigram ($\tau = 0.8$)</td>
<td>0.63</td>
<td>0.97</td>
<td>0.46</td>
</tr>
<tr>
<td>4-Shingling ($\tau = 0.3$)</td>
<td>0.81</td>
<td>0.89</td>
<td>0.73</td>
</tr>
<tr>
<td>I-Match (IDF=[0.0, 0.8])</td>
<td>0.50</td>
<td>0.53</td>
<td>0.47</td>
</tr>
<tr>
<td>SpotSigs (#A=500, $\tau = 0.4$)</td>
<td>0.70</td>
<td>0.77</td>
<td>0.64</td>
</tr>
</tbody>
</table>
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
  - $C(A) = 4$

Cohesive Campaign Extraction

A B C D E F G H

unvisited
neighbors
visiting
visited

heap

Clique Size vs Vertices

1 2 3 4
Start from A, explore A’s neighbor B.
Calculate $C(a) = 2$ and output it.
Cohesive Campaign Extraction

![Graph with nodes A, B, C, D, E, F, G, H and clique sizes](image)

- unvisited
- neighbors
- visiting
- visited

![Bar chart showing clique sizes with vertex A](image)
Cohesive Campaign Extraction

Mark A visited. From B, explore B’s immediate neighbors CFH. Calculate $CC(A,B) = 2$ and output it.
Mark A visited. From B, explore B’s immediate neighbors CFH. Calculate \( CC(A,B) = 2 \) and output it.
Cohesive Campaign Extraction

The diagram illustrates a network of vertices labeled A through H, with connections between some of them. The vertices are color-coded as follows:

- **unvisited**: light grey
- **neighbors**: yellow
- **visiting**: red
- **visited**: dark grey

A heap is shown on the right, containing vertices C, F, and H.

A bar graph is also present, showing clique sizes for vertices AB, with clique sizes of 2 and 4.
Cohesive Campaign Extraction

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.
Mark B visited. Choose C as next visiting vertex. From C, explore C’s immediate neighbors DFGH. Calculate $\text{CC}(B,C) = 4$ and output it.
Cohesive Campaign Extraction

The curve represents a cohesive campaign.

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

<table>
<thead>
<tr>
<th>Approach</th>
<th>NumC</th>
<th>$F_1$</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loose</td>
<td>12</td>
<td>0.962</td>
<td>0.986</td>
<td>0.940</td>
</tr>
<tr>
<td>Strict</td>
<td>12</td>
<td>0.906</td>
<td>0.907</td>
<td>0.904</td>
</tr>
<tr>
<td>Cohesive</td>
<td>11</td>
<td>0.963</td>
<td>0.977</td>
<td>0.950</td>
</tr>
<tr>
<td>$k$-means</td>
<td>5</td>
<td>0.89</td>
<td>1</td>
<td>0.805</td>
</tr>
</tbody>
</table>
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

![Campaigns graph]

- spam
- promotion
- template
- celebrity
- babble

- top 50
- top 100
- top 200
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

<table>
<thead>
<tr>
<th>Msgs</th>
<th>Users</th>
<th>Talking Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>560</td>
<td>34</td>
<td>Iron Man 2 spam</td>
</tr>
<tr>
<td>401</td>
<td>390</td>
<td>Facebook photo template</td>
</tr>
<tr>
<td>231</td>
<td>231</td>
<td>Support Breast Cancer Research (short link)</td>
</tr>
<tr>
<td>218</td>
<td>218</td>
<td>Formspring template</td>
</tr>
<tr>
<td>203</td>
<td>197</td>
<td>Chat template (w/ link)</td>
</tr>
<tr>
<td>166</td>
<td>166</td>
<td>Support Breast Cancer Research (full link)</td>
</tr>
<tr>
<td>165</td>
<td>154</td>
<td>Quote “send to anyone u don’t regret meeting”</td>
</tr>
<tr>
<td>153</td>
<td>153</td>
<td>Justin Bieber Retweets</td>
</tr>
<tr>
<td>145</td>
<td>31</td>
<td>Twilight Movie spam</td>
</tr>
<tr>
<td>111</td>
<td>111</td>
<td>Quote “This October has 5 Fridays ...”</td>
</tr>
</tbody>
</table>
User Level Campaign Detection

User ID | User Messages
--- | ---
1 | M1: Support Breast Cancer Awareness, add a #twibbon
2 | M1: Support Breast Cancer Awareness, add a #twibbon
3 | M1: I'm having fun with @formspring. Create an account
   M2: follow me at http://formspring.me/xnadjeeaaa
4 | M1: @Wookiefoot Real Money Doubling Forex Robot Fap
5 | M1: @justinbieber Support Breast Cancer Awareness, add
6 | M1: RT @justinbieber Support … #twibbon to
User Level Campaign Detection

62 campaigns (>= 4 users)

28 campaigns (>= 4 users)

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


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

1. Get Info
2. Feed Misinformation

[2] Campaigns (Groups of users)

[1] Social Spam (Individual Spammer/Content Polluter)

3. Attention

Origin: Crowdturfing

Popular keywords or topics

Keywords or topics

Social Spam (Individual Spammer/Content Polluter)
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.

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.

<table>
<thead>
<tr>
<th>Case</th>
<th># of unique tweets</th>
<th>% of re-tweets</th>
<th># of unique “affirms”</th>
<th># of unique “denies”</th>
<th># of unique “questions”</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Confirmed truths</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The international airport of Santiago is closed</td>
<td>301</td>
<td>81</td>
<td>291</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>The Viña del Mar International Song Festival is canceled</td>
<td>261</td>
<td>57</td>
<td>256</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Fire in the Chemistry Faculty at the University of Concepción</td>
<td>42</td>
<td>49</td>
<td>38</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Navy acknowledges mistake informing about tsunami warning</td>
<td>135</td>
<td>30</td>
<td>124</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Small aircraft with six people crashes near Concepción</td>
<td>129</td>
<td>82</td>
<td>125</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Looting of supermarket in Concepción</td>
<td>160</td>
<td>44</td>
<td>149</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Tsunami in Iloca and Duao towns</td>
<td>153</td>
<td>32</td>
<td>140</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>TOTAL</td>
<td>1181</td>
<td>1123</td>
<td>4</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>AVERAGE</td>
<td>168.71</td>
<td>160.43</td>
<td>0.57</td>
<td>4.29</td>
<td></td>
</tr>
<tr>
<td><strong>False rumors</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Death of artist Ricardo Arjona</td>
<td>50</td>
<td>37</td>
<td>24</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>Tsunami warning in Valparaiso</td>
<td>700</td>
<td>4</td>
<td>45</td>
<td>605</td>
<td>27</td>
</tr>
<tr>
<td>Large water tower broken in Rancagua</td>
<td>126</td>
<td>43</td>
<td>62</td>
<td>38</td>
<td>20</td>
</tr>
<tr>
<td>Cousin of football player Gary Medel is a victim</td>
<td>94</td>
<td>4</td>
<td>44</td>
<td>34</td>
<td>2</td>
</tr>
<tr>
<td>Looting in some districts in Santiago</td>
<td>250</td>
<td>37</td>
<td>218</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>“Huascar” vessel missing in Talcahuano</td>
<td>234</td>
<td>36</td>
<td>54</td>
<td>66</td>
<td>63</td>
</tr>
<tr>
<td>Villarrica volcano has become active</td>
<td>228</td>
<td>21</td>
<td>55</td>
<td>79</td>
<td>76</td>
</tr>
<tr>
<td>TOTAL</td>
<td>1682</td>
<td>502</td>
<td>836</td>
<td>216</td>
<td></td>
</tr>
<tr>
<td>AVERAGE</td>
<td>240.29</td>
<td>71.71</td>
<td>119.43</td>
<td>30.86</td>
<td></td>
</tr>
</tbody>
</table>
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

- \( \geq 5 \) agreement
Spreading a specific news/event

OR

Conversation or comments among friends.
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
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.

<table>
<thead>
<tr>
<th>Class</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Prec.</th>
<th>Recall</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEWS</td>
<td>0.927</td>
<td>0.039</td>
<td>0.922</td>
<td>0.927</td>
<td>0.924</td>
</tr>
<tr>
<td>CHAT</td>
<td>0.874</td>
<td>0.054</td>
<td>0.892</td>
<td>0.874</td>
<td>0.883</td>
</tr>
<tr>
<td>UNSURE</td>
<td>0.873</td>
<td>0.07</td>
<td>0.86</td>
<td>0.873</td>
<td>0.866</td>
</tr>
<tr>
<td>W. Avg.</td>
<td>0.891</td>
<td>0.054</td>
<td>0.891</td>
<td>0.891</td>
<td>0.891</td>
</tr>
</tbody>
</table>

89% accuracy

## Table 7: Results for the credibility classification.

<table>
<thead>
<tr>
<th>Class</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Prec.</th>
<th>Recall</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (“true”)</td>
<td>0.825</td>
<td>0.108</td>
<td>0.874</td>
<td>0.825</td>
<td>0.849</td>
</tr>
<tr>
<td>B (“false”)</td>
<td>0.892</td>
<td>0.175</td>
<td>0.849</td>
<td>0.892</td>
<td>0.87</td>
</tr>
<tr>
<td>W. Avg.</td>
<td>0.860</td>
<td>0.143</td>
<td>0.861</td>
<td>0.860</td>
<td>0.86</td>
</tr>
</tbody>
</table>

86% accuracy
Detecting fake images on Twitter during Hurricane Sandy
Background: Hurricane Sandy

- Category 3 storm
- Damages worth $75 billion USD
- Coast of NE America [Atlantic ocean]

Motivation

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
Goal and Methodology

• Goal: Detecting tweets containing fake images

• Methodology
Data Description – Total Sandy Dataset

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Tweets</td>
<td>1,782,526</td>
</tr>
<tr>
<td>Total unique users</td>
<td>1,174,266</td>
</tr>
<tr>
<td>Tweets with URLs</td>
<td>622,860</td>
</tr>
</tbody>
</table>
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

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Tweets with fake images</td>
<td>10,350</td>
</tr>
<tr>
<td>Users with fake images</td>
<td>10,215</td>
</tr>
<tr>
<td>Tweets with real images</td>
<td>5,767</td>
</tr>
<tr>
<td>Users with real images</td>
<td>5,678</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>User Features [F1]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Friends</td>
</tr>
<tr>
<td>Number of Followers</td>
</tr>
<tr>
<td>Follower-Friend Ratio</td>
</tr>
<tr>
<td>Number of times listed</td>
</tr>
<tr>
<td>User has a URL</td>
</tr>
<tr>
<td>User is a verified user</td>
</tr>
<tr>
<td>Age of user account</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tweet Features [F2]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length of Tweet</td>
</tr>
<tr>
<td>Number of Words</td>
</tr>
<tr>
<td>Contains Question Mark?</td>
</tr>
<tr>
<td>Contains Exclamation Mark?</td>
</tr>
<tr>
<td>Number of Question Marks</td>
</tr>
<tr>
<td>Number of Exclamation Marks</td>
</tr>
<tr>
<td>Contains Happy Emoticon</td>
</tr>
<tr>
<td>Contains Sad Emoticon</td>
</tr>
<tr>
<td>Contains First Order Pronoun</td>
</tr>
<tr>
<td>Contains Second Order Pronoun</td>
</tr>
<tr>
<td>Contains Third Order Pronoun</td>
</tr>
<tr>
<td>Number of uppercase characters</td>
</tr>
<tr>
<td>Number of negative sentiment words</td>
</tr>
<tr>
<td>Number of positive sentiment words</td>
</tr>
<tr>
<td>Number of mentions</td>
</tr>
<tr>
<td>Number of hashtags</td>
</tr>
<tr>
<td>Number of URLs</td>
</tr>
<tr>
<td>Retweet count</td>
</tr>
</tbody>
</table>
Classification Results

<table>
<thead>
<tr>
<th></th>
<th>F1 (user)</th>
<th>F2 (tweet)</th>
<th>F1+F2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>56.32%</td>
<td>91.97%</td>
<td>91.52%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>53.24%</td>
<td>97.65%</td>
<td>96.65%</td>
</tr>
</tbody>
</table>

• 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

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>14:00 ~ 14:10</td>
<td>Introduction to Social Media Threats (Social Spam, Campaigns, Misinformation and Crowdturfing)</td>
</tr>
<tr>
<td>14:10 ~ 14:55</td>
<td>Social Spam</td>
</tr>
<tr>
<td>14:55 ~ 15:30</td>
<td>Campaigns</td>
</tr>
<tr>
<td>15:30 ~ 16:00</td>
<td>Break</td>
</tr>
<tr>
<td>16:00 ~ 16:30</td>
<td>Misinformation</td>
</tr>
<tr>
<td>16:30 ~ 17:10</td>
<td>Crowdturfing</td>
</tr>
<tr>
<td>17:10 ~ 17:30</td>
<td>Challenges, Tools and Conclusion</td>
</tr>
</tbody>
</table>
Conceptual Level of Tutorial Theme

[1] Social Spam (Individual Spammer/Content Polluter)

[2] Campaigns (Groups of users)


1. Get Info
2. Feed Misinformation
3. Attention
4. Origin: Crowdturfing

Popular keywords or topics
**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 an interesting task](#) > [Work](#) > [Earn money](#)

[Find HITs Now](#)

**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

[Find your account] > [Load your tasks] > [Get results]

[Get Started](#)

[Learn more about being a Worker](#)
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?

---

**Real-time Crowd Labor**

4 judgments/sec
Current velocity

911,585,246
Total judgments

**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 skills 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]

<table>
<thead>
<tr>
<th>Website</th>
<th>Campaigns’ % Crowd-turfing</th>
<th>Tasks</th>
<th>$ per Subm.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon Turk (US)</td>
<td>12%</td>
<td>2.9M</td>
<td>$0.092</td>
</tr>
<tr>
<td>ShortTask* (US)</td>
<td>95%</td>
<td>527K</td>
<td>$0.096</td>
</tr>
<tr>
<td>MinuteWorkers (US)</td>
<td>70%</td>
<td>10K</td>
<td>$0.241</td>
</tr>
<tr>
<td>MyEasyTask (US)</td>
<td>83%</td>
<td>4K</td>
<td>$0.149</td>
</tr>
<tr>
<td>Microworkers (US)</td>
<td>89%</td>
<td>84K</td>
<td>$0.175</td>
</tr>
</tbody>
</table>

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

Crowdturfing Workflow

Customers
- Initiate campaigns
- May be legitimate businesses

Agents
- Manage campaigns and workers
- Verify completed tasks

Workers
- Complete tasks for money
- Control Sybils on other websites

Company X
ZBJ/SDH
Worker Y
Campaign Information

Promote our product using your blog

<table>
<thead>
<tr>
<th>Campaign ID</th>
<th>Promote our product using your blog</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input Money</td>
<td>¥100 元</td>
</tr>
<tr>
<td>Category</td>
<td>Blog Promotion</td>
</tr>
<tr>
<td>Rewards</td>
<td>100 tasks, each ¥0.8, 77 submissions accepted, Still need 23 more</td>
</tr>
<tr>
<td>Status</td>
<td>Ongoing (177 reports submitted)</td>
</tr>
</tbody>
</table>

Report generated by workers

- Report ID: 2814244号
- WorkerID: WYQ591456
- Experience: 10 中校
- Reputation: 4 钻石

URL: http://a8356ab.blog.163.com/blog/sr?report_id=2814244

Accepted! Screenshots:

- Screen 1
- Screen 2
- Screen 3
### High Level Statistics

<table>
<thead>
<tr>
<th>Site</th>
<th>Active Since</th>
<th>Total Campaigns</th>
<th>Workers</th>
<th>Reports</th>
<th>$ for Workers</th>
<th>$ for Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZBJ</td>
<td>Nov. 2006</td>
<td>76K</td>
<td>169K</td>
<td>6.3M</td>
<td>$2.4M</td>
<td>$595K</td>
</tr>
</tbody>
</table>

#### Site Growth Over Time

- **Campaigns per Month**: ZBJ, SDH
- **Dollars per Month**: $
Spam Per Worker

- Transient workers
  - Makes up majority of a diverse worker population

- Prolific workers
  - Major force of spam generation
Are Workers Real People?

The graph shows the percentage of reports from workers across different times of the day. The x-axis represents hours in the day, ranging from 0 to 20. The y-axis represents the percentage of reports from workers, ranging from 0 to 9.

Key times of the day are marked on the graph:
- Late Night/Early Morning
- Work Day/Evening
- Lunch
- Dinner

The graph uses two curves to represent different data sets:
- ZBJ
- SDH

The graph illustrates how the percentage of reports varies throughout the day, with peaks during certain times and troughs at others.
## Campaign Types

### Top 5 Campaign Types on ZBJ

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

- 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?

• 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 category
4. Vote for TOMMY MARSH and BAD DOG

Top Rated

- Tommy Marsh & Bad Dog
  - 320 votes

- Don 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

<table>
<thead>
<tr>
<th>Class</th>
<th>User Profiles</th>
<th>Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workers</td>
<td>2,864</td>
<td>364,581</td>
</tr>
<tr>
<td>Non-Workers</td>
<td>9,878</td>
<td>1,878,434</td>
</tr>
</tbody>
</table>
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

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

• 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:**

<table>
<thead>
<tr>
<th>Class</th>
<th>User Profiles</th>
<th>Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workers</td>
<td>2,864</td>
<td>364,581</td>
</tr>
<tr>
<td>Non-Workers</td>
<td>9,878</td>
<td>1,878,434</td>
</tr>
</tbody>
</table>

• **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)**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Workers</th>
<th>Non-workers</th>
</tr>
</thead>
<tbody>
<tr>
<td>links in tweets /</td>
<td>tweets</td>
<td>0.696</td>
</tr>
<tr>
<td>tweets / recent days</td>
<td>4</td>
<td>37</td>
</tr>
<tr>
<td>@username in tweets / recent days</td>
<td>2</td>
<td>28</td>
</tr>
<tr>
<td>the number of posted tweets per day</td>
<td>3</td>
<td>21</td>
</tr>
<tr>
<td>rt in tweets /</td>
<td>tweets</td>
<td>0.7</td>
</tr>
<tr>
<td>Swearing in LIWC</td>
<td>0.001</td>
<td>0.009</td>
</tr>
<tr>
<td>links in RT tweets /</td>
<td>RT tweets</td>
<td>0.589</td>
</tr>
<tr>
<td>Anger in LIWC</td>
<td>0.003</td>
<td>0.012</td>
</tr>
<tr>
<td>Total Pronouns in LIWC</td>
<td>0.054</td>
<td>0.107</td>
</tr>
<tr>
<td>1st Person Singular in LIWC</td>
<td>0.019</td>
<td>0.051</td>
</tr>
</tbody>
</table>
Detecting Crowd Workers (Cont’d)

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

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>F1</th>
<th>AUC</th>
<th>FNR</th>
<th>FPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>93.26%</td>
<td>0.966</td>
<td>0.955</td>
<td>0.036</td>
<td>0.174</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Class</th>
<th>User Profiles</th>
<th>Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workers</td>
<td>368</td>
<td>40,344</td>
</tr>
</tbody>
</table>

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

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
  – 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**
  - Pig: [https://pig.apache.org/](https://pig.apache.org/)
  - Hive: [https://hive.apache.org/](https://hive.apache.org/)
  - Cascalog: [http://cascalog.org/](http://cascalog.org/)
  - Giraph: [https://giraph.apache.org/](https://giraph.apache.org/)

- **Scalable machine learning**
  - Mahout: [https://mahout.apache.org/](https://mahout.apache.org/)

- **Large scale stream processing**
  - Summingbird: [https://github.com/twitter/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:
All Reference List

- Grier, C., Thomas, K., Paxson, V., and Zhang, M. @spam: the underground on 140 characters or less. In CCS, 2010.
- Lee, K., Eoff, B., and Caverlee, J. Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter. In ICWSM, 2011.
All Reference List

Thanks to…

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