Detecting Collective Attention Spam

Kyumin Lee, James Caverlee, Krishna Y. Kamath, Zhiyuan Cheng

Texas A&M University

WebQuality :: April 16, 2012
Example of Collective Attention (1)
Example of Collective Attention (1)

- The video theme of the 2016 Olympic Games on YouTube
Example of Collective Attention (2)

- The bursts of popularity of events on Twitter

  ![Tweeting Big Events](image)

  - **New Year’s Eve (midnight in Japan)**
  - **Obama’s Speech on Osama Bin Laden**: 5,000 tweets per second in the peak and a sustained average rate of 3,000 tweets per second over several hours
  - **Super Bowl 2011**
  - **Japan Over Denmark During the World Cup**
  - **2010 MTV Video Music Awards**
  - **The Royal Wedding**
  - **Oscars 2011**
  - **Obama Inauguration**
Viral Marketing and Collective Attention

Trends

Google Trends
Hot Searches (USA)
1. etch a sketch
2. chipper jones
3. trayvon martin
4. lim lebow jets
5. sean payton
6. new york jets
7. saints bounty program
8. reese witherspoon pregnant
9. ny jets
10. hunger games

Twitter Trends
United States trends
Beliebers Are Very Sexy  #msnNOW  Klout Bomb  #tedxbyu
Matt Forte  Angry Birds Space  Etch A Sketch  Chipper
Collective Attention Spam

- Definition: spammers target social media where user attention quickly coalesces and then collectively focuses around a phenomenon.
Amanda Knox was accused of the murder of Meredith Kercher in Italy.

She served four years of a 26-year sentence before the murder conviction was overturned on October 3, 2011.

Amanda Knox Freed After Appeal in Italian Court

Amanda Knox in a Perugia, Italy, courtroom on Monday after her conviction was overturned. More Photos »

By ELISABETTA POVOLEDO
Published: October 3, 2011
Collective Attention Spam Example (1)

Amanda Knox: Murder on Trial VIDEO

dowopor | 12 videos |  

Click on the Link In The Description To Watch This Video

Click on [http://hotznews.com/wow/](http://hotznews.com/wow/) - FREE to watch Amanda Knox: Murder on
Collective Attention Spam Example (2)

Results for #DrakeCriesWhen

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

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

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

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

MaryaLccketto9
Omg...Is this real?? bit.ly/oFalmo #gritosmexicanos #DrakeCriesWhen #FastFoodAddiction Glen Rice
16 seconds ago

ClarethaPrevot3
bit.ly/oFalmo #gritosmexicanos #DrakeCriesWhen #FastFoodAddiction Glen Rice
16 seconds ago
In this work

• Examine properties of collective attention spam in Twitter trending topics
  – Longevity of spam accounts
  – The total amount of collective attention spam
  – The properties of accounts engaged in such behavior

• Develop a machine learning based spam classifier to detect spam messages generated by collective attention spammers
  – Evaluate over a small collection of trending topics
Collective Attention Spammers and Their Tactics

- Dataset: 5.3 million messages containing top trending topics on Twitter (3.7% messages are spam)

While we were collecting the messages and the users, we periodically checked whether the users were suspended or not.
Properties of Spam Accounts (1)

- Properties of Spammers (left) and Regular Users (right)

<table>
<thead>
<tr>
<th></th>
<th># of Followings</th>
<th># of Followers</th>
<th># of Messages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Max</td>
<td>67,579</td>
<td>189,805</td>
<td>228,634</td>
</tr>
<tr>
<td>Avg</td>
<td>104</td>
<td>183</td>
<td>567</td>
</tr>
<tr>
<td>Median</td>
<td>0</td>
<td>0</td>
<td>9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th># of Followings</th>
<th># of Followers</th>
<th># of Messages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Max</td>
<td>298,287</td>
<td>5,211,919</td>
<td>610,869</td>
</tr>
<tr>
<td>Avg</td>
<td>324</td>
<td>506</td>
<td>4,984</td>
</tr>
<tr>
<td>Median</td>
<td>156</td>
<td>116</td>
<td>1,625</td>
</tr>
</tbody>
</table>

- Longevity of Spammers

Longevity = suspended date – creation date

There are three peaks.
How long does it take for Twitter safety team to suspend spam accounts?

The time = suspended time – collected time
74% of spammers were suspended within 1 day.
5% living over four days (out-of-range of the figure displayed).
A large widow of opportunity for polluting the collective attention.

Word cloud of spam messages
Spammers often used celebrity names, such as Angelina Jolie and Justin Bieber.
They also used some adult / provocative terms such as “naked”, “hot” and “sexy”.

Properties of Spam Accounts (3)

• URLs in spam messages
  – Multiple accounts posted either the same URL or variant URL.
  – Ex) “bit.ly/oJ4AZn” + “?=” + “random code”

Check https://bitly.com/oJ4AZn+
There are several peaks!
Detecting Collective Attention Spam: A Case Study

• Goal
  – predict whether a message containing a trending topic is a spam message through a classifier as soon as possible

• Evaluation Metrics
  – Accuracy
  – False positive rate
  – False negative rate
  – Total spam detection (TSD)

\[ TSD_{\text{topic}}(\%) = \frac{\text{detected spam messages}}{\text{total # of spam messages in the topic}} \]
Data and Features

• Data
  – Messages associated with seven popular trending topics in the dataset.

• Main features
  – # of URLs
  – # of hashtags
  – # of @mentions
  – Is a message retweeted?
  – The length of a message
  – The length of a payload (after removing @mentions, URLs and hashtags)

• The original features + bag-of-words features
  – The main features
  – Each term in a message as a feature
Experimental Results (1)

- #DearHair

[Graphs showing non-spam and spam messages, total spam detection percent, accuracy, and false positive rate over hours.]
Experimental Results (2)

- The other six trending topics with the main features

<table>
<thead>
<tr>
<th>Topic</th>
<th>Training Time</th>
<th>TSD (%)</th>
<th>Accuracy</th>
<th>FP</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td>#TheyNeedToBringBack</td>
<td>First 4 hours</td>
<td>74.42%</td>
<td>99.47%</td>
<td>0.002</td>
<td>0.148</td>
</tr>
<tr>
<td>#WhatYouShouldKnowAboutMe</td>
<td>First 3 hours</td>
<td>60.25%</td>
<td>99.08%</td>
<td>0.002</td>
<td>0.347</td>
</tr>
<tr>
<td>#MeAndYouCantDate</td>
<td>First 5 hours</td>
<td>51.28%</td>
<td>98.79%</td>
<td>0.001</td>
<td>0.409</td>
</tr>
<tr>
<td>#IfICouldDoItOverAgain</td>
<td>First 3 hours</td>
<td>68.25%</td>
<td>99.02%</td>
<td>0.001</td>
<td>0.307</td>
</tr>
<tr>
<td>#YouNeedToRealize</td>
<td>First 1 hour</td>
<td>80.11%</td>
<td>98.92%</td>
<td>0.002</td>
<td>0.172</td>
</tr>
<tr>
<td>#YouKnowBetter</td>
<td>First 3 hours</td>
<td>71.49%</td>
<td>97.67%</td>
<td>0.002</td>
<td>0.272</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>First 3 hours</td>
<td>68%</td>
<td>98.83%</td>
<td>0.001</td>
<td>0.275</td>
</tr>
</tbody>
</table>

- **Summary**
  - It is possible to detect and prevent collective attention spam messages by learning early-age spam messages in a topic.
  - Achieved a **high TSD** and **accuracy** and **low false positive rate**.
  - Open question: how to verify that the spam messages in the **first few hours** used to bootstrap the learning approach are indeed spam.
Conclusions and Future Work

• Examined the problem of collective attention spam, studied the presence of it on Twitter, and tested an initial approach for early detection.

• Learning a classifier based on the first moments of a bursting phenomenon is effective to detect spam messages in the future as more attention focuses.

• Future work
  – Expanding our study to a larger number of trending topics to better understand the robustness of the proposed approach.
  – Investigate adaptations of the developed methods for alternative domains (e.g., YouTube videos).
Detecting Collective Attention Spam

kyumin@cse.tamu.edu
students.cse.tamu.edu/kyumin
Twitter: @humanist0810