Identification and Analysis of Malicious Content on Facebook: A Survey

Prateek Dewan (PhD1111)
Comprehensive Examination
November 24, 2014

Committee Members
Dr. Sambuddho Chakravarty (IIITD)
Dr. Anupam Joshi (IIITD / UMBC)
Dr. Anand Kashyap (Symantec Research)
Dr. Ponnurangam K. (Advisor)
Outline

• What is Facebook?
• Why Facebook?
• Malicious content on Facebook
• Identification techniques
• Research gaps
• Opportunities
What is Facebook?

• World’s biggest Online Social Network (OSN)
• Users connect by mutual consent
• Used to
  • keep in touch with friends and family
  • share what people are up to
  • consume information about real world events

Why Facebook?

• Largest online social network in the world
  • 1.32 billion monthly active users
  • 4.75 billion posts per day
  • Over 300 petabytes of data

• Spammers exploit context of event to lure victims into scams.

• Facebook spammers make $200 million just by posting links.²

Example (1 / 2)
Example (2 / 2)
Types of malicious content on Facebook

**ADVERTISING**
- Apple Inspired Car Charger Plug Adapter and Lightning Data Cable for iPhone 5 I eBay [Link]

**SCAMS**
- People were asked to sign a petition in support of the player Luis Suárez used as bait for Facebook scam [Link]

**FAKE INFORMATION**
- Facebook pages set up in the names of flight MH17 victims link to porn sites and malware [Link]
Focus

Online Social Media

Malicious Content

Events

Gao et al., IMC 2010, Detecting and characterizing social spam campaigns

Rahman et al., USENIX 2012, Efficient and scalable software detection in online social networks

Thomas et al., IMC 2011, Suspended accounts in retrospect: An analysis of Twitter spam

Benevenuto et al., SIGIR 2009, Detecting spammers and content promoters in online video social networks

Vieweg et al., CHI 2010, Microblogging during two natural hazard events

Hughes et al., ISCRAM 2009, Twitter adoption and use in mass convergence and emergency events

Mendoza et al., SOMA 2010, Twitter under crisis: Can we trust what we RT?

Gupta et al., PSOSM 2012, Credibility ranking of tweets during high impact events
Focus

Online Social Media

Malicious Content

Events

- Vieweg et al., CHI 2010, Microblogging during two natural hazard events
- Hughes et al., ISCRAM 2009, Twitter adoption and use in mass convergence and emergency events
- Mendoza et al., SOMA 2010, Twitter under crisis: Can we trust what we RT?
- Gupta et al., PSOSM 2012, Credibility ranking of tweets during high impact events
- Gao et al., IMC 2010, Detecting and characterizing social spam campaigns
- Rahman et al., USENIX 2012, Efficient and scalable software detection in online social networks
- Thomas et al., IMC 2011, Suspended accounts in retrospect: An analysis of Twitter spam
- Benevenuto et al., SIGIR 2009, Detecting spammers and content promoters in online video social networks
Facebook’s efforts to counter malicious content

• Facebook Immune System, Stein et al., 2011


• Two billion dollar lawsuit against fake “Likes” spam, 2014
Research efforts to combat malicious content on Facebook
Overview

I

Data

1. Posts containing URLs
2. User profiles

II

Collection Techniques

1. Snowball sampling
2. Convenient sampling
3. Honeypots

III

Identification Techniques

1. Unsupervised learning
2. Supervised learning

IV

Evaluation metrics

1. Precision / Recall
2. True positives
3. False negatives
4. Purity / Inverse purity
5. Manual verification
Overview

I

Data
1. Posts containing URLs
2. User profiles

II

Collection Techniques
1. Snowball sampling
2. Convenient sampling
3. Honeypots

III

Identification Techniques
1. Unsupervised learning
2. Supervised learning

IV

Evaluation metrics
1. Precision / Recall
2. True positives
3. False negatives
4. Purity / Inverse purity
5. Manual verification
Identification techniques

• Unsupervised learning
  • Clustering based on message similarity
  • Markov Clustering (MCL)
Identification techniques

• Unsupervised learning
  • Clustering based on message similarity
  • Markov Clustering (MCL)
Clustering based on message similarity

• Data - 187 million wall posts
  • 2.08 million wall posts containing a URL

• Collection technique - Snowball sampling

• Methodology
  • Post <description, URL>
  • Same URL / similar description → cluster
  • “Distributed” coverage + “Bursty” nature = spam cluster
    • distributed nature: > 5 users in a cluster
    • bursty nature: median time between posts < 90 minutes

Detecting and Characterizing Social Spam Campaigns, Gao et al., IMC 2010
Clustering based on message similarity

• Results
  • 1.4 million clusters in total
  • 297 clusters (212k posts) satisfied distributed and bursty thresholds

• Evaluation
  • 93.9% true positives; verified manually

• Highlights
  • Largest dataset in literature
  • False negative estimation missing

Detecting and Characterizing Social Spam Campaigns, Gao et al., IMC 2010
Identification techniques

• Unsupervised learning
  • Clustering based on message similarity
  • Markov Clustering (MCL)
Markov clustering

• Data - 320 user profiles
  • 165 spammers, 155 legitimate

• Collection technique - Convenient sampling

• Methodology
  • Modelled social networks as weighted graph $G = (V, E, W)$
  $$W(E_{ij}) = |F_{ij}^a| + |P_{ij}| + |U_{ij}|$$
  common active friends common page likes fraction of common URLs

• Applied Markov Clustering (MCL)

An MCL-based Approach for Spam Profile Detection in Online Social Networks, Ahmed et al., IEEE TrustCom 2012
Markov clustering

• Results
  • $F_p = 0.88$ (Harmonic mean of purity and inverse purity)
  • $F_B = 0.79$ (Harmonic mean of precision and recall)

• Highlights
  • Technique uses only 3 features, yet achieves good results
  • Small dataset; could be biased
Summary

I
Data
1. Posts containing URLs
2. User profiles

II
Collection Techniques
1. Snowball sampling
2. Convenient sampling
3. Honeypots

III
Identification Techniques
1. Unsupervised learning
2. Supervised learning

IV
Evaluation metrics
1. Precision / Recall
2. True positives
3. False negatives
4. Purity / Inverse purity
5. Manual verification
Identification techniques

- Supervised learning
  - Support Vector Machine (SVM)
  - Decision Trees
  - Random Forests
Identification techniques

• Supervised learning
  • Support Vector Machine (SVM)
  • Decision Trees
  • Random Forests
Support Vector Machine

• Data - 7,500 wall posts containing URLs
  • 2,500 malicious, 5,000 legitimate

• Collection technique - Convenient sampling

• Methodology
Support Vector Machine

• Results - 60k posts marked malicious out of 40M posts
  • True positive rate: 97% (manually verified)
  • False positive rate: 0.005% (manually verified)

• Highlights
  • Work aimed to detect “socware”
  • Real world deployment - MyPageKeeper
  • Classifier heavily relies on bag-of-words and message similarity
Identification techniques

• Supervised learning
  • Support Vector Machine (SVM)
  • Decision Trees
  • Random Forests
Decision trees

- Data - 187 million wall posts
  - 2.08 million wall posts containing a URL

- Collection technique - Snowball sampling

- Methodology

Towards Online Spam Filtering in Social Networks, Gao et al., NDSS 2012
Decision trees

• Results
  • True positive rate: 80.9%
  • False positive rate: 0.19%
  • Throughput: 1,580 messages / second

• Highlights
  • Model stays accurate even after 9 months of training
  • Model tested on Twitter data as well
  • All “new” instances marked as legitimate; system cannot detect campaigns initially outside its learning

Towards Online Spam Filtering in Social Networks, Gao et al., NDSS 2012
Identification techniques

• Supervised learning
  • Support Vector Machine (SVM)
  • Decision Trees
  • Random Forests
Random forests

• Data - 3,831 user profiles
  • 173 spammers, 827 legitimate profiles used for training

• Collection technique - Honeypots

• Methodology
  • Six features extracted from profiles in training set
  • Applied Random forest classifier on feature vectors
Random forests

• Results - 10 fold cross validation on training set
  • False positive rate: 2%
  • False negative rate: 1%
  • 130 more profiles marked as spam from 790k profiles in test set (7 false positives)

• Highlights
  • Only work to use honeypot approach
  • Approach tested on Twitter too
  • Features used not available publicly (FF ratio, friend choice...)

Detecting Spammers on Social Networks, Stringhini et al., ACSAC 2010
Summary

I

Data
1. Posts containing URLs
2. User profiles

II

Collection Techniques
1. Snowball sampling
2. Convenient sampling
3. Honeypots

III

Identification Techniques
1. Unsupervised learning
2. Supervised learning

IV

Evaluation metrics
1. Precision / Recall
2. True positives
3. False negatives
4. Purity / Inverse purity
5. Manual verification
# In a nutshell

<table>
<thead>
<tr>
<th></th>
<th>User detection</th>
<th>Post detection</th>
<th>Real time</th>
<th>Real world deployment</th>
<th>Validation on other OSNs</th>
<th>Profile features</th>
<th>Content features</th>
<th>Network features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gao et al., 2010, IMC</strong></td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td><strong>Stringhini et al., ACSAC 2010</strong></td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td><strong>Gao et al., 2012, NDSS</strong></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td><strong>Rahman et al., 2012 USENIX</strong></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td><strong>Ahmed et al., 2012, TrustCom</strong></td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>
Other work on malicious content detection

• Facebook
  - Malicious URL detection on Facebook, Guan et al., JWIS 2011
  - FRAppE: Detecting malicious applications on Facebook, Rahman et al., CoNEXT 2012

• Other OSNs
  - Suspended accounts in retrospect: An analysis of Twitter spam, Thomas et al., IMC 2011
  - Detecting spammers and content promoters in online video social networks, Benevenuto et al., SIGIR 2009
  - Uncovering social spammers: social honeypots + machine learning, Lee et al., SIGIR 2010
Challenges in conducting research on Facebook

• Approximately 28% users share their data with an audience wider than their friends. ³

• Public data is not rich
  • Network features not accessible publicly
  • Limited user profile features available

• No real time API endpoint to fetch data in real time

³ http://www.consumerreports.org/cro/magazine/2012/06/facebook-your-privacy/index.htm
Straight from Facebook

• Interaction with Christopher Palow, Engineering Manager, Facebook (fb.com/palow)

  • “We’re not really fond of people who crawl Facebook for research.”

  • “Facebook’s filters don’t catch all the bad content. If something goes undetected in real time, it is usually detected within the next 24 hours. If not, it’s gone forever.”

  • “Web of Trust is mostly reputation based whereas SURBL is spam honey pots. We actually have access to both but we don’t keep our integrations well updated.”

• Sample bash script to look for spammers

  ```bash
grep -P -i -o '_wau.push\(\"small\", \"S+\", \" /tmp/bad_site.html | cut -d , -f 2 | cut -d "" -f 2 | awk '{print "http://whos.amung.us/stats/history/" $1 "}'}'
  ```
Research gaps

- Large scale studies for detecting malicious content
  - Ignore posts without URLs
  - Work on detecting malicious posts which are part of a campaign
  - Rely partially on passive features (likes, comments etc.) which take time to build up; not suitable for real time detection
  - Don’t rely on profile features
- Crawled Facebook - ethical implications
  - Did so prior to 2009
Research gaps

• Existing techniques used to detect malicious content on other social networks cannot be directly ported to Facebook
  • Lack of publicly available information
  • Difference in network dynamics and usage
Opportunities

• Efficient, real time, self learning and evolving techniques to detect malicious content

• Study content without URLs

• Light weight client-side solutions for malicious content identification
Thanks!

prateekd@iiitd.ac.in
http://precog.iiitd.edu.in/people/prateek/
cerc.iiitd.ac.in
Backup slides
Bayes classification

• Methodology

• Collected 49,110 posts on top 20 Facebook pages in July 2010 (to Feb. 2011).
  • 5,637 malicious; 43,473 benign
  • Posts made by users on page; NOT the posts made by page itself.

• Filtered posts containing URLs, looked for URLs blacklisted by PhishTank, SURBL, WOT, Google SafeBrowsing etc.

• Extracted 7 features (4 URL, 3 message containing the URL)

• Used Bayes classification model to classify URLs as malicious / benign
Bayes classification

• Results
  • Authors report an accuracy of 94.9% (TPR: 95%, FNR: 3.56%)

• Validation
  • Technique validated using top 20 Facebook pages in Europe and Asia
  • Achieved approx. 90% accuracy and TPR in both; performance felt due to skewed malicious:benign ratio in datasets
Malicious apps

• FRAppE (Facebook’s Rigorous Application Evaluator)
  • Identified 6,273 malicious apps (from MyPageKeeper) and 6,273 benign apps (from Social Bakers)
  • Characterised app behaviour; extracted on-demand and aggregation-based features

• Observations
  • Malicious apps redirect users to domains with poor reputation
  • 97% of malicious apps request for only “publish” permissions
Malicious apps

• Results
  • FRAppE Lite
    • Extracts on-demand features - Accuracy: 99%, FN rate: 4.4%
  • FRAppE
    • Extracts on-demand and aggregation-based features - Accuracy: 99.5%, FN rate: 4.1%

• Validation
  • 98.5% apps validated as malicious (apps deleted, app name similarity, typo squatting, posted link similarity etc.)