Smoke Screener or Straight Shooter: Detecting Elite Sybil Attacks in User-Review Social Networks

Haizhong Zheng\textsuperscript{1} Minhui Xue\textsuperscript{2} Hao Lu\textsuperscript{1} Shuang Hao\textsuperscript{3}
Haojin Zhu\textsuperscript{1} Xiaohui Liang\textsuperscript{4} Keith Ross\textsuperscript{2}

\textsuperscript{1}Shanghai Jiao Tong University
\textsuperscript{2}New York University Shanghai
\textsuperscript{3}University of Texas at Dallas
\textsuperscript{4}University of Massachusetts Boston

February, 2018
User-Review Social Networks

Yelp: Search restaurants in San Diego
Reviews in User-Review Social Networks (URSNs)
Sybil Attacks in URSNs

MarketWatch: Yelp deems 20% of user reviews ‘suspicious’
Published: Sept 27, 2013 8:36 a.m. ET

Financial Times: TripAdvisor fined in Italy over false reviews
Online travel website to challenge €500,000 penalty

Phoenix Finance: Dianping deleted 6 million fake reviews in half a year.
Examples of fake reviews

Which one is the fake review?

Review 1

Review 2
Examples of fake reviews

A fake review posted by the volunteer

A real review posted by my friend
A New Type of Sybils

Elite Sybil users consist of two kinds of users.

1. Sybil accounts creating reviews not belonging to Sybil tasks to mimic benign users.
2. Accounts owned by benign users, but occasionally fulfilling a Sybil task to get the payments.

Elite Sybil User

Regular Sybil User
Outlines

01 Elite Sybil Attacks
02 ELISEDET
03 Evaluation & Measurement
04 Conclusion
Customers (or Overhyped stores)
Stores want to boost their scores rapidly.

Agents
Agents are responsible for accepting tasks from customers and launching campaigns.

Leaders
Leaders take charge of recruiting workers and distributing tasks to workers.

Elite Sybil workers
Elite Sybil workers are internet users post fake reviews for profit.
01 Customers (or Overhyped stores)
Stores want to boost their scores rapidly.

02 Agents
Agents are responsible for accepting tasks from customers and launching campaigns.

03 Leaders
Leaders take charge of recruiting workers and distributing tasks to workers.

04 Elite Sybil workers
Elite Sybil workers are internet users post fake reviews for profit.
Customers (or Overhyped stores)
Stores want to boost their scores rapidly.

Agents
Agents are responsible for accepting tasks from customers and launching campaigns.

Leaders
Leaders take charge of recruiting workers and distributing tasks to workers.

Elite Sybil workers
Elite Sybil workers are internet users post fake reviews for profit.
**ELITE SYBIL ATTACKS**

**Customers (or Overhyped stores)**
Stores want to boost their scores rapidly.

**Agents**
Agents are responsible for accepting tasks from customers and launching campaigns.

**Leaders**
Leaders take charge of recruiting workers and distributing tasks to workers.

**Elite Sybil workers**
Elite Sybil workers are internet users post fake reviews for profit.

**Sybil Organization Architecture**

[Diagram showing the organization structure with Customers, Agents, Leaders, and Elite Sybil workers.]
Customers (or Overhyped stores)
Stores want to boost their scores rapidly.

Agents
Agents are responsible for accepting tasks from customers and launching campaigns.

Leaders
Leaders take charge of recruiting workers and distributing tasks to workers.

Elite Sybil workers
Elite Sybil workers are internet users post fake reviews for profit.
01 ELITE SYBIL ATTACKS

Leader-supervised model
The reviews generated by elite Sybil workers

Leader hands-on model
The content of reviews is provided by leaders.

The quality of fake reviews is strictly controlled!!!
Why do Sybil organizations put so much effort to control the quality of reviews?
The payments vary with ratings of accounts:

<table>
<thead>
<tr>
<th>Ratings of Accounts</th>
<th>Rewards per Submission</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-star, 1-star</td>
<td>$0.30</td>
</tr>
<tr>
<td>2-star</td>
<td>$0.75</td>
</tr>
<tr>
<td>3-star</td>
<td>$1.50</td>
</tr>
<tr>
<td>4-star</td>
<td>$3.74</td>
</tr>
<tr>
<td>5-star, 6-star</td>
<td><strong>$5.98</strong></td>
</tr>
</tbody>
</table>
ELITE SYBIL ATTACKS

Challenges

Weak connectivity of User-Review Social Networks (URSNs)

Graph-based detection approaches may not work.

Highly adaptive and professional

Imitate real users in user profiles and review content. Behavior-based detection approaches may not work.

More active out of Sybil campaigns

Lower percentage of fake reviews in their posts. Clustering approaches may not work.
Outlines

01 Elite Sybil Attacks
02 ELISEDET
03 Evaluation & Measurement
04 Conclusion
02 ELISEDET

Sybil Community Detection

Campaign Window Detection

Elite Sybil User Detection
Sybil Community Detection

Campaign Window Detection

Elite Sybil User Detection
Challenge 1: Weak connectivity
Challenge 2: Professional fake reviews

Sybil Community Detection
Campaign Window Detection
Elite Sybil User Detection

Challenge 1: Build the relation between users via collusive reviews.
Challenge 1: Weak connectivity

Challenge 2: Professional fake reviews

Sybil Community Detection

Campaign Window Detection

Elite Sybil User Detection

Challenge 2: Use features irrelevant to user profiles and review content.
(community-based features, network features, content-free user features)
ELISEDET

Sybil Community Detection

Campaign Window Detection

Elite Sybil User Detection

Overview

Data → Sybil Community Detection
    → Construct Graph
    → Louvain Method
    → Classifier

Undetected User → Sybil Community

Campaign Window Detection

Elite Sybil User Detection
    → Participation Rate
    → Sybilness Score
    → Elite Sybil User
Sybil Community Detection

Campaign Window Detection

Elite Sybil User Detection

Challenge 3: More active out of Sybil campaign

Challenge 3: Use campaign windows to filter out the reviews irrelevant to Sybil Campaigns
Collusive reviews

Posted by two users in the same store in a short time period $\Delta T$, with the same rating (same goal).

Collusive relation between users

\[ R(U_1, U_2) = \frac{|\text{Collusive Reviews}|}{|\text{Reviews of } U_1| + |\text{Reviews of } U_2|} \]

\[ R(U_1, U_2) = \frac{5}{5 + 5} = 0.5 \]
Detect Sybil communities in three steps:

1. Build Relations between Users
   - Build collusive relations and construct graphs

2. Louvain Community Detection
   - Detect communities

3. Classification
   - Classify click farming communities

Sybil Community-based Features
Community-based Features
Network Features
Content-free User Features
Sparse interval:
The number of weeks with at least one review is less than the number of weeks without any reviews.

Iteratively find and delete sparse review intervals

Algorithm 1: Detecting Campaign Time Windows

Input: A list $L_{\text{review}}$ whose item $L_{\text{review}}[i]$ denotes the number of reviews posted in the $i$th week.
Output: The start point $l$ and end point $r$ of the campaign time window.

1: $l \leftarrow 0$;
2: $r \leftarrow \text{length}(L_{\text{review}}) - 1$;
3: while (true) do
4:  $I_{l,l'} \leftarrow \text{find(left, } l\) \{\text{Find the first sparse interval } I_{l,l'} \text{ from left.}\}$
5:  $I_{r',r} \leftarrow \text{find(right, } r\) \{\text{Find the first sparse interval } I_{r',r} \text{ from right.}\}$
6:  if ($l' = r$ and $r' = l$) then \{There is no sparse interval.\}
7:    break;
8:  end if
9:  if ($|I_{l,l'}| \leq |I_{r',r}|$) then \{Choose the interval with fewer reviews.\}
10:     $l \leftarrow l' + 1$;
11:  else
12:     $r \leftarrow r' - 1$;
13:  end if
14: end while
15: return $l, r$;
Participation rate between users and communities

Just consider the reviews posted in Sybil campaigns.

\( N_C(k) \): the accumulated number of reviews posted in the kth time window of community C.

\( N_C^{max} \): the maximum number of reviews posted within all time windows of community C.

\[
P_C(k) = \frac{N_C(k)}{N_C^{max}}
\]

Indicate the importance of kth time interval

\[
N_{u\in C} = \sum_k P_C(k) \cdot N_{u\in C}(k)
\]

Indicate the importance of reviews u posted for community C

Participation rate: \( \rho_{u\in C} = \frac{1}{1 + \exp\left(-\frac{N_{u\in C} - \mu_C}{\sigma_C}\right)} \), for any \( u \in C \), Standardize with sigmoid function

Sybilness

\[
f(u) = \sum_C \rho_{u\in C} \cdot N_{u\in C}
\]

Perceived likelihood indicating if a user is an elite Sybil user
EVALUATION

Detection Results

Dataset
32,933 stores, 3,555,154 users and 10,541,931 reviews

Community
566 Sybil communities with 22,324 users and 144 benign communities with 5,222 users

Elite Sybil User
1. User $u$ does not belong to any Sybil communities.
2. There is a community $C$ and the user participation rate $\rho_{u \in C}$ is larger than 0.5 (average participation rate in a community).
3. 12,292 elite Sybil users.
Detection Results

Dataset

32,933 stores, 3,555,154 users and 10,541,931 reviews

Community

566 Sybil communities with 22,324 users and 144 benign communities with 5,222 users

Elite Sybil User

1. User $u$ does not belong to any Sybil communities.
2. There is a community $C$ and the user participation rate $\rho_{u \in C}$ is larger than 0.5 (average participation rate in a community).
3. 12,292 elite Sybil users.

<table>
<thead>
<tr>
<th>Type</th>
<th># Stores</th>
<th># Overhyped Stores</th>
<th>Percentage of Overhyped Stores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cinema</td>
<td>235</td>
<td>71</td>
<td>30.21%</td>
</tr>
<tr>
<td>Hotel</td>
<td>1,738</td>
<td>134</td>
<td>7.71%</td>
</tr>
<tr>
<td>Restaurant</td>
<td>22,474</td>
<td>1,244</td>
<td>5.54%</td>
</tr>
<tr>
<td>Entertainment</td>
<td>1,384</td>
<td>73</td>
<td>5.27%</td>
</tr>
<tr>
<td>Wedding Service</td>
<td>320</td>
<td>8</td>
<td>2.50%</td>
</tr>
<tr>
<td>Beauty Store</td>
<td>1,460</td>
<td>35</td>
<td>2.40%</td>
</tr>
</tbody>
</table>
EVALUATION

Performance

Sybil Community

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision tree</td>
<td>93.80 %</td>
<td>92.90 %</td>
<td>93.60 %</td>
<td>92.83 %</td>
</tr>
<tr>
<td>SVM</td>
<td>96.74 %</td>
<td>96.47 %</td>
<td>96.45 %</td>
<td>99.42 %</td>
</tr>
<tr>
<td>GNB</td>
<td>94.21 %</td>
<td>93.44 %</td>
<td>93.57 %</td>
<td>97.64 %</td>
</tr>
<tr>
<td>KNN</td>
<td>96.75 %</td>
<td>96.47 %</td>
<td>96.50 %</td>
<td>97.45 %</td>
</tr>
<tr>
<td>Ada boost</td>
<td>93.84 %</td>
<td>93.54 %</td>
<td>93.60 %</td>
<td>97.92 %</td>
</tr>
<tr>
<td>Random forest</td>
<td>93.16 %</td>
<td>94.01 %</td>
<td>92.99 %</td>
<td>97.42 %</td>
</tr>
</tbody>
</table>

Elite Sybil User

<table>
<thead>
<tr>
<th>Type of users</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sybilness top 1,000 users</td>
<td>93.80%</td>
</tr>
<tr>
<td>Random 1,000 users</td>
<td>90.70%</td>
</tr>
</tbody>
</table>
User-level star ratings

Compared to regular Sybil users, elite Sybil users have higher user levels.
Some Sybil communities provide services for a huge number of overhyped stores.
Relation between Communities and Stores

1. Except for red dots, other dots with same colors are chain stores.
2. 12.37% of Sybil communities post fake reviews for chain stores.

- Sybil communities
- Overhyped stores

Chain store
User posting period

Elite Sybil users post fake reviews periodically. Evading detection method basing on posting frequency.
Three different Sybil communities posted fake reviews in a store and caused the review spikes. Increase in the number of reviews.
Case Study

Increase in rating

The reviews spikes lead to increases in the rating of the store.
CONCLUSION

Novel attack: Elite Sybil attack
Elite Sybil users are more spread out temporally, craft better edited contents, but have fewer reviews filtered

*Elite Sybil Detection System: ELISEDET*
Both highly effective and scalable as a standalone system

Measurement about elite Sybil users
Elite Sybil have higher user-levels and deliberately control the posting time
Thanks for listening!

Questions?

sjtu.zhenghaizhong@gmail.com