Turning the Tide: Curbing Deceptive Yelp Behaviors

SIAM SDM 2014

June 2018 @ UCAS

Dongwon Lee

Penn State / IST
Review Centric Social Networks

User → Reviews → Social networks

Reviews

User

Social networks

Example reviews from Yelp and Foursquare.
Malicious Behaviors: Fraudulent Reviews

- Social networks: ideal targets for malicious behaviors
- Up to 25% of Yelp reviews are fraudulent [1]

**YELP:**

- Extra half-star rating causes a restaurant to sell out 19% more often [2]
- One-star increase leads to a 5–9% increase in revenue [3]

Malicious Behaviors: Review Campaigns

- Review Campaign: post multiple fraudulent reviews
- Example: Search Engine Optimization (SEO) companies [1]
  - Use IP spoofing techniques
  - Set up fake online profiles
  - Target Yelp, Google Local, CitySearch
  - Investigated by law enforcement

- Deceptive venue: uses review campaigns to alter rating

Feasibility Study

- Created fake venues
- Posted review jobs on Amazon Mechanical Turk
- Received more than 90 (fake) reviews

... for a fistful of dollars
Problem Statement

- Detect malicious behaviors in review centric social networks
  - Fraudulent reviews
  - Deceptive venues
  - Impactful review campaigns
Adversary Model

- Needs to adjust the rating of target venue
- Has finite budget
- Controls finite set of (IP address, Yelp Sybil account) pairs
- Has access to a market of review writers
- Social network provider does not collude with attackers
Marco System Overview

1. Friend & review count
2. Venue “expertise”
3. Venue activities
4. ...

Deceptive & legitimate venues

RSD Module

ARD Module

FRI Module

Venue Classifier

Features

Venue timeline

Review ratings

7,435 venues
195,417 users
270,121 reviews

Users
Reviews/Time
Venues

Friend relations

Spatial vicinity

Fraudulent & genuine reviews

Train

Label
Successful Review Campaign

- Increases (decreases) the rating of the target venue by at least half a star

- **Claim:** The minimum number of reviews the adversary needs to post in order to fraudulently increase the rating of a venue by half a star is $n/7$
  
  - $n$: the number of genuine reviews a venue has at the completion of the campaign
Review Spikes

**Theorem:** If $n > 49$, a successful review campaign will exceed, during the attack interval, the maximum number of reviews of a uniform review distribution.
Review Spike Detection (RSD)

- Identify venues that receive higher number of positive (negative) reviews than normal

- Use the measures of dispersion of Box-and-Whisker plots to detect outliers

- Two features
  - Number of spikes detected for a venue
  - Normalized amplitude of the highest spike
Marco System Overview

1. Friend & review count
2. Venue “expertise”
3. Venue activities
4. ...

7,435 venues
195,417 users
270,121 reviews

Deceptive & legitimate venues

RSD Module
ARD Module
FRI Module

Venue timeline
Review ratings

Review Classifier

Fraudulent & genuine reviews
**Aggregate Rating Disparity (ARD) Module**

- **ARD Module:** Measure the review *divergence*
  - **N:** total number of reviews of venue V

\[
ARD(V) = \frac{\sum_{1}^{N} |\text{Review rating} - \text{avg rating of V}|}{N}
\]

Graph showing the comparison between venue average rating and user rating over days.
Marco System Overview

1. Friend & review count
2. Venue “expertise”
3. Venue activities
4. …

Deceptive & legitimate venues

FRI Module
- Review Classifier
  1. Friend & review count
  2. Venue “expertise”
  3. Venue activities
  4. …

ARD Module - Review ratings

RSD Module - Venue timeline

Venue Classifier

Features

Users - Reviews/Time - Venues

Friend relations

Spatial vicinity

7,435 venues
195,417 users
270,121 reviews
Fraudulent Review Impact (FRI) Module

- Venues with few genuine reviews
  - Vulnerable to review campaigns
  - Long term campaigns can re-define the “normal” review posting behavior

- **FRI Module**: detect fraudulent reviews that significantly impact the aggregate rating of venues
Features to classify review (fraudulent vs. genuine):
- Review writer
  - Number of friends
  - Number of reviews written
  - Expertise of user around venue
- Number of check-ins at venue
- Number of photos at venue
- Age of user’s account when review was posted
- Feedback count of review

FRI Feature:
- Percentage of reviews classified as fraudulent
Review Data

- Gold standard fraudulent reviews
  - Spelp (spam Yelp) sites
  - Suspicious user accounts
  - Generic review text

- Gold standard genuine reviews
  - Written by active, popular users
  - No short, generic reviews

- 200 fraudulent and 202 genuine reviews
Review Classification

Overall accuracy: RF [94%], Bagging [93.5%], DT [93%]
Marco System Overview

1. Friend & review count
2. Venue "expertise"
3. Venue activities
4. ...

Features

Venue Classifier

Deceptive & legitimate venues

RSD Module
Venue timeline

ARD Module

Review ratings

FRI Module
Review Classifier

Users

Reviews/Time

Venues

Friend relations

Spatial vicinity

Fraudulent & genuine reviews

7,435 venues
195,417 users
270,121 reviews

Train

Label

Train

Label

7,435 venues
195,417 users
270,121 reviews

Train
Venue Classification Features

Features to classify venues:

- Number of review spikes for venue
- Amplitude of the highest spike
- Aggregate rating disparity
- Fraudulent review impact of venue
- Count of reviews classified fraudulent
- Rating of the venue
- Number of reviews (with check-ins & photos)
- Age of venue

RSD
ARD
FRI
Venue Data

- **Deceptive venue:** fraudulent reviews impact its rating

- **Ground truth:** Yelp’s “Consumer Alert” feature
Venue Data (cont’d)

- Gold standard legitimate venues
  - Well known consistent quality
  - At most 10% of reviews are filtered by Yelp

- 90 deceptive and 100 legitimate venues
RF and DT are tied for best accuracy, 95.8%.
Comparison with state-of-the-art

Compare Marco with the three deceptive venue detection strategies of Feng et al. [1], \(\text{avg}\Delta\), \(\text{dist}\Phi\) and peak↑.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marco/RF</td>
<td>95.8</td>
</tr>
<tr>
<td>(\text{avg}\Delta)</td>
<td>66.3</td>
</tr>
<tr>
<td>(\text{dist}\Phi)</td>
<td>72.1</td>
</tr>
<tr>
<td>peak↑</td>
<td>58.9</td>
</tr>
</tbody>
</table>

Per-module overhead

Zoom-in of FRI module overhead
Marco in the Wild: Yelp Data

- **YCrawl**: developed crawler to fetch raw HTML pages of Yelp venue and user accounts

- Collected:
  - 7,435 venues from San Francisco, New York City and Miami
    - Car shops, Spas, Moving companies
  - 270,121 reviews
  - 195,417 reviewers
Experimental Results on Live Data

San Francisco: Marco flags almost 10% of car repair and moving companies as suspicious

<table>
<thead>
<tr>
<th>City</th>
<th>Car Shop</th>
<th>Mover</th>
<th>Spa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Miami, FL</td>
<td>1000 (6)</td>
<td>348 (8)</td>
<td>1000 (21)</td>
</tr>
<tr>
<td>San Francisco, CA</td>
<td>612 (59)</td>
<td>475 (45)</td>
<td>1000 (42)</td>
</tr>
<tr>
<td>NYC, NY</td>
<td>1000 (8)</td>
<td>1000 (27)</td>
<td>1000 (28)</td>
</tr>
</tbody>
</table>

Detected deceptive venues by Marco out of collected venues in Yelp
Conclusions

- Lower bound on the number of reviews required to launch *successful* review campaign
- Marco: automatic detection of fraudulent reviews, deceptive venues and impactful review campaigns
- Novel dataset of reviews and venues
- Marco is effective and fast