Malicious Behavior on the Web: Characterization and Detection

Srijan Kumar (@srijankr)
Justin Cheng (@jcccf)
Jure Leskovec (@jure)

Slides are available at http://snap.stanford.edu/www2017tutorial/
Tutorial Outline

Malicious users
- Trolling
- Sockpuppets
- Vandals

Misinformation
- Fake reviews
- Hoaxes

Web: Source of information
"Don't believe everything you read on the Internet"
- Abraham Lincoln
Types of false information

Misinformation
honest mistake

Disinformation
deliberate lie to mislead
Reviews

Google Play

Amazon

Yelp

4.0

5 stars: 530
4 stars: 157
3 stars: 96
2 stars: 42
1 star: 101

928 total reviews

Helpfulness

All Versions
Impact of Fake Reviews

Makhija et al., 2016, Luca et al., 2011

+1 increase in star rating increases revenue by 5-9%

Makhija et al., 2016, Luca et al., 2011
Characteristics of fake reviews and reviewers
STRONG DECEPTIVE INDICATORS

A focus on who they were with
In this example, “My husband;” also words like “family.”

Greater use of first-person singular
Fake reviews tend to use “I” and “me” more often.

Direct mention of where they stayed
Hotel and city names were less common in truthful reviews, which focus more on details about the hotel itself, like “small” or “bathroom.”

“My husband and I stayed in the [hotel name] Chicago and had a very nice stay! The rooms were large and comfortable. The view of Lake Michigan from our room was gorgeous. Room service was really good and quick, eating in the room looking at that view, awesome! The pool was really nice but we didn’t get a chance to use it. Great location for all of the downtown Chicago attractions such as theaters and museums. Very friendly staff and knowledgable, you can’t go wrong staying here.”

SLIGHT DECEPTIVE INDICATORS

High adverb use
“Very” and “really” are both used twice; “here” is used once.

High verb use
“Get”, “go”, “use”, “can’t”, “didn’t”, “eating”, “had”, “looking”, “stayed”, “was” (three times), “were.”

Use of “I” and positive emotion
Deceptive reviews tend to use exclamation points, while truthful reviews used more punctuation of other kinds, including “$.”
Fake reviewers are more opinionated

Kumar et al., 2017
Fake reviewers

Fake reviewers give fewer reviews
Fake reviewers write shorter reviews

1. Fake reviewers give fewer reviews
2. Fake reviewers write shorter reviews

Mukherjee et al., 2013
Fake reviewers are faster and have bimodal rating pattern

Kumar et al., 2017, Li et al., 2017
Fake reviewers collude

Kumar et al., 2017
Detecting fake reviewers
• User is suspicious if his behavior deviates substantially from that of the global model
• Global Model:
  • Users belong to different cluster, each representing a different behavior
  • Each cluster is associated with a common Dirichlet prior, to model the common behavior of users in the cluster
  • The property is drawn using a multinomial derived from the cluster’s Dirichlet prior
Each user has a multinomial rating distribution vector, drawn from a cluster-specific Dirichlet prior.
BIRDNEST

Cluster 1

Time difference distributions

Cluster 2

Time difference distributions

Hooi et al., SDM 2016
BIRDNEST Results

Precision at $k$ of BIRD (Flipkart)

100% precision @ top 50

AWESOMEApp4FreeMoney!!! $$$$$$

All first time users will need a CODE after downloading this app. So download it now and use my CODE for bonus points. CODE: ...
Intuition: Fair reviewers upvote and fake reviewers downvote good products. Fair reviewers downvote bad products and fake reviewers upvote bad products.

Unsupervised Loopy Belief Propagation algorithm

Add behavior property: include a prior to indicate its suspiciousness

Use cumulative distribution of the property over all users

\[
f(x_{li}) = \begin{cases} 
1 - P(X_l \leq x_{li}), & \text{if high is suspicious (H)} \\
P(X_l \leq x_{li}), & \text{otherwise (L)} 
\end{cases}
\]

\[
S_i = 1 - \sqrt{\frac{\sum_{l=1}^{F} f(x_{li})^2}{F}}
\]

Rayana et al., KDD 2015
SpEagle Results

Behavior is more important than text, but it still helps.
Iterative algorithm to compute 3 interdependent measures:

**Trustworthiness of reviewer** which depends (non-linearly) on its reviews’ honesty scores;

**Reliability of store** depending on the trustworthiness of the reviewers writing reviews for it and the score;

**Honesty of review** which is a function of reliability of the store and trustworthiness of store reviewers.
Iteratively calculate three interdependent metrics:

**Fairness of each user** who writes a review: how fair is the user in giving correct reviews?

**Reliability of each review**: how trustworthy is each review itself?

**Goodness of each product**: what is the quality of the product?
FairJudge

Fairness
F(u) [0,1]

Reliability
R(u,p) [0,1]

Goodness
G(p) [-1,1]

Kumar et al., 2017
Fairness

\[ F(u) = \frac{\sum_{(u,p) \in \text{Out}(u)} R(u,p)}{|\text{Out}(u)|} \]

Kumar et al., 2017
Goodness

\[ G(p) = \frac{\sum_{(u,p) \in \text{In}(p)} R(u,p) \cdot \text{score}(u,p)}{|\text{In}(p)|} \]

Kumar et al., 2017
How far is the rating from the goodness of product

$$R(u, p) = \frac{1}{2} (F(u) + (1 - \frac{|\text{score}(u, p) - G(p)|}{2}))$$

How fair is the user who gives the rating

Kumar et al., 2017
Initialization

F(u) = 1

F(u) = 1

F(u) = 1

R(u,p) = 1

R(u,p) = 1

R(u,p) = 1

G(p) = 1

G(p) = 1

G(p) = 1

F(u) = 1

F(u) = 1

F(u) = 1

F(u) = 1
Updating Goodness - Iteration 1

\[
G(p) = \frac{\sum_{(u, p) \in \text{In}(p)} R(u, p) \cdot \text{score}(u, p)}{|\text{In}(p)|}
\]

\[F(u) = 1\]

\[R(r) = 1\]

\[G(p) = 0.67\]

\[R(r) = 1\]

\[F(u) = 1\]

\[G(p) = 0.67\]

\[R(r) = 1\]

\[G(p) = -0.67\]

\[R(r) = 1\]

\[F(u) = 1\]
Updating Reliability - Iteration 1

\[ R(u, p) = \frac{1}{2}(F(u) + (1 - \frac{|\text{score}(u, p) - G(p)|}{2})) \]

- \( R(r) = 0.92 \)
- \( F(u) = 1 \)
- \( R(r) = 0.92 \)
- \( G(p) = 0.67 \)
- \( G(p) = 0.67 \)
- \( G(p) = -0.67 \)
- \( F(u) = 1 \)
- \( F(u) = 1 \)
Updating Fairness - Iteration 1

\[ F(u) = \frac{\sum_{(u,p) \in \text{Out}(u)} R(u,p)}{|\text{Out}(u)|} \]

- **F(u)** = 0.92
- **R(r)** = 0.92
- **G(p)** = 0.67
- **R(r)** = 0.92
- **F(u)** = 0.92
- **R(r)** = 0.92
- **G(p)** = -0.67
- **R(r)** = 0.92
- **F(u)** = 0.92
- **R(r)** = 0.92
- **F(u)** = 0.92
- **R(r)** = 0.92
FairJudge - After convergence

F(u) = 0.17
R(r) = 0.17
G(p) = 0.67
F(u) = 0.83
R(r) = 0.83
G(p) = -0.67
F(u) = 0.83
R(r) = 0.83
G(p) = 0.67
F(u) = 0.83
R(r) = 0.83
G(p) = 0.67
F(u) = 0.83
R(r) = 0.83
G(p) = 0.67
F(u) = 0.83
Cold Start Problem

Most reviewers give few ratings and most products receive few ratings.

Solution: add Bayesian priors

\[ F(u) = \frac{0.5 \cdot \alpha + \sum_{(u,p) \in \text{Out}(u)} R(u, p)}{\alpha + |\text{Out}(u)|} \]

\[ G(p) = \frac{\sum_{(u,p) \in \text{In}(p)} R(u, p) \cdot \text{score}(u,p)}{\beta + |\text{In}(p)|} \]

Kumar et al., 2017
Incorporating Behavioral Properties

Rating distribution

Timestamp distribution

Use BIRDNEST score of reviewers and products
FairJudge

\[
F(u) = \frac{0.5 \cdot \alpha_1 + \alpha_2 \cdot IBIRDNEST_{IRT_D} (u) + \sum_{(u, p) \in \text{out}(u)} R(u, p)}{\alpha_1 + \alpha_2 + |\text{out}(u)|}
\]

\[
R(u, p) = \frac{1}{2} (F(u) + (1 - \frac{|\text{score}(u, p) - G(p)|}{2}))
\]

\[
G(p) = \frac{\beta_2 \cdot IBIRDNEST_{IRT_D} (p) + \sum_{(u, p) \in \text{In}(p)} R(u, p) \cdot \text{score}(u, p)}{\beta_1 + \beta_2 + |\text{In}(p)|}
\]

Time complexity \(O(k|E|)\)

\(k\) is the number of iterations, which is bounded. \(|E|\) is the number of edges.

Kumar et al., 2017
Detecting Fair Reviewers

Kumar et al., 2017
Detecting Fake Reviewers

80 of 100 reported fake reviewers in Flipkart correct. 
FairJudge is in use at Flipkart.

Kumar et al., 2017
Importance of components

N = Network
C = Cold Start Solution
B = Behavior

Kumar et al., 2017
Summary: Fake Reviewers

- **Fake Reviewers**: Users who write non-truthful reviews for products
- Fake reviews are worse: shorter, more positive, use more “I”s and more verbs and adverbs
- Fake reviewers are deceptive: they collude among themselves and are faster
- Textual, behavioral and network based algorithms can detect fake reviewers
- Combination of several components performs the best
References

S. Kumar, B. Hooi, D. Makhija, M. Kumar, C. Faloutsos and V.S. Subrahmanian. FairJudge: Trustworthy User Prediction in Rating Platforms. arXiv 1703.10545


A. Mukherjee, V. Venkataraman, B. Liu, and N. S. Glance. What yelp fake review filter might be doing? In ICWSM, 2013.

References


A. Mishra and A. Bhattacharya. Finding the bias and prestige of nodes in networks based on trust scores. In WWW, 2011.

Additional slides
Fake reviews are more positive

Yoo et al., 2009
Lemma: \[ |G^\infty(p) - G^1(p)| \leq 1 \]

Error bound:
The error between iterations is bounded, and as \( t \) increases, the rating scores converge. The error bound is given by:

\[ |F^\infty(u) - F^t(u)| \leq \frac{3^t}{4} \]

\[ |R^\infty(r) - R^t(r)| \leq \frac{3^t}{4} \]

\[ |G^\infty(p) - G^t(p)| \leq 3^{(t-1)} \]

As \( t \) increases, \( F^t(u) \to F^\infty(u), G^t(p) \to G^\infty(p), R^t(r) \to R^\infty(r) \)

Kumar et al., 2017