Fraud Detection in Online Reviews

Mudit Mehrotra Rajendra Kumar Raghupatruni
mmehrotra@cs.stonybrook.edu rraghupatrun@cs.stonybrook.edu
Department of Computer Science, Stony Brook University

Abstract

Reviews about products by their users are helpful to the prospective customers in taking a constructive decision. However, often spammers post fake online reviews in order to promote or demote a product or business. It is a challenge for the automated machines to identify fake reviews. Several techniques have been developed to identify such reviews and spammer groups. One such technique is the FraudEagle framework that exploits network effects among reviewers and products for opinion fraud detection. In this report, we discuss an extension to the FraudEagle framework by adding new attributes to the review network such as *helpfulness*, *verified purchase* and *duplicates in reviews* for improved opinion spam detection.

Keywords – Review, fraud, spam.

1. INTRODUCTION

The problem of fraud detection is about identifying fake reviews from a pool of reviews posted online. This is a challenging problem as spammers post reviews that appear benign but are actually untruthful. Sometimes they even post genuine reviews for camouflage. There are experienced individuals and groups who publish untruthful reviews to make money. As such, it is a challenge for automated fake opinion detecting systems to identify such reviews.

The aim of this project is to improve the existing Fraud eagle technique by incorporating new orthogonalities into the review network graph. The new attributes added to the review network (esp., to the User Nodes) are *helpfulness*, *verified purchase* and *duplicates in reviews*.

2. PRIOR WORK

Past research in the field of identifying fake reviews has been multifaceted. Pattern mining and spam indicator heuristics have been exploited by Mukherjee et al. (Mukherjee, Liu, Glance & Jindal, 2011) to study group spamming behaviors. Each reviewer is associated with a group using frequent pattern mining to form candidate groups. These groups are then ranked on the basis of indicators such as content similarity across group, group size, group support, time window and whether the group posted early review in order to make a big impact. We are trying to use the group content similarity which can be an indicator of presence of duplicate reviews.

Wang et al. (Wang, Xie and Yu, 2011) proposed use of a heterogeneous review graph to understand the interrelationship between trustiness of reviewers, honesty of reviews and the reliability of

stores. Their technique uses connectivity structure of reviewer's reviews, all the stores he/she reviewed and reviews from other reviewers. This paper is the first graph based approach to detect online spam. However, in our approach we used network effects as proposed in FraudEagle.

Network effect among reviewers and products have been exploited by Akoglu et al. (Akoglu, Chandy & Faloutsos, 2013) for opinion fraud detection. The proposed algorithm is unsupervised and scalable and is applicable to large datasets. This is simple and powerful algorithm to identify spam based on sentiment of reviews with network effects included. However, the attributes used in the graph are limited only to review rating. Our approach is to add more attributes such as *helpfulness*, *verified purchase* and *duplicates in review* to improve the existing algorithm.

3. DATA COLLECTION

The dataset of both recommended and non-recommended reviews is obtained from Yelp.com, we used an automatic crawler to get the details (like Restaurant Name, Author, Review, Rating, Votes for Useful/Cool/Funny, Timestamp etc.) of the reviews on the restaurants. We have also collected the academic dataset available directly from the website: https://www.yelp.com/academic_dataset

We also manually crawled the data from Amazon to obtain few reviews which have "Verified Purchased" tags which we intend to use as an attribute for the network graph nodes.

4. BACKGROUND

In this project, we are extending the existing FraudEagle framework for improved opinion spam detection by adding attributes such as *helpfulness*, *verified purchase* and *duplicates in reviews*. FraudEagle uses users and products to form the nodes of a bipartite graph. Each node can be classified as {honest, fraud} for a user, {good, bad} for a product. It utilizes **pairwise Markov Random Fields (pMRF)** for this classification. pMRF is a set of random variables that satisfy the pairwise Markov property described by an undirected graph i.e. any two non adjacent variables are conditionally independent given all other variables.

Labels, $\mathcal{L}^{\mathcal{U}}$ = {Honest, Dishonest} represents the domain of users and $\mathcal{L}^{\mathcal{P}}$ = {Good, Bad} represents the domain of products.
G(V, E), a signed review network graph in which users and products form the nodes. They are connected with signed links {+,-}.
Let $\Psi_i^U \in \psi$ be a prior mapping $\Psi_i^U : \mathcal{L}^{\mathcal{U}} \rightarrow R_{\geq 0}$, $\Psi_j^P : \mathcal{L}^{\mathcal{P}} \rightarrow R_{\geq 0}$ for each unobserved user Y_i and unobserved product Y_j
For each edge, let $\Psi_{ij}^S \in \psi$ be a compatibility mapping.

The probability of users/products belonging to each class (as per pMRF) is given by

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{Y_i \in \mathcal{Y}^{\mathcal{V}}} \psi_i(y_i) \prod_{e(Y_i^{\mathcal{U}}, Y_j^{\mathcal{P}}, s) \in \mathcal{E}} \psi_{ij}^s(y_i, y_j)$$

The probability assigned to the unobserved variables might not be their best assignment. The best assignment is determined by **Loopy Belief Propagation** (LBP) extended to work on a signed network.

LBP is a message passing algorithm for inferring classification in a graphical model. In this, each user Y_i and product Y_j sends a message to its neighbour indicating how it thinks the neighbour should be classified. Message passing continues until all the messages stabilize (no change). The objective of the problem is to maximize P(y/x), defined in the above equation.

$$m_{i\to j}(y_j) = \alpha_1 \sum_{y_i \in \mathcal{L}^{\mathcal{U}}} \psi_{ij}^{s}(y_i, y_j)\, \psi_i^{\mathcal{U}}(y_i) \prod_{Y_k \in \mathcal{N}_i \cap \mathcal{Y}^{\mathcal{P}} \setminus Y_j} m_{k\to i}(y_i), \quad \forall y_j \in \mathcal{L}^{\mathcal{P}}$$

$$b_i(y_i) = \alpha_2\, \psi_i^{\mathcal{U}}(y_i) \prod_{Y_j \in \mathcal{N}_i \cap \mathcal{Y}^{\mathcal{P}}} m_{j\to i}(y_i), \forall y_i \in \mathcal{L}^{\mathcal{U}}$$

Where $m_{i \to j}$ is a message sent by user Y_i to product Y_j, α's are the normalization constants. $\mathcal{L}^{\mathcal{U}}$ denotes the label domain for users and $\mathcal{L}^{\mathcal{P}}$ denotes the label domain for products. b_i (yi) of assigning Y_i with label y_i

5. ALGORITHM

We are using the Fraud Eagle framework as a baseline for this project, the idea is to improve the results by considering few more attributes (orthogonalities) like helpfulness of the review, duplicates in the review text and verified purchased tags. Helpfulness in the review (available in the Yelp dataset) has three types Useful, Funny and Cool, each of this is associated with the count value indicating that some user(s) who wish to avail the service or purchase the product found this review useful. Out of these only the useful count is taken into consideration for this work, as Funny and Cool may not actually convey the review usefulness. Duplicate reviews are compared using the cosine similarity. Based on this each user who wrote duplicate reviews and the review for a product are given a score to indicate the untruthfulness of the review and reviewer. Verified Purchase tag is the endorsement given by Amazon (in amazon dataset) to indicate that the user purchased the product from amazon. If the purchase is verified then the review is more likely to be genuine. With these attributes we are trying to improve the existing framework.

Outline

Step 1: Scoring

signedInferenceAlgorithm()

Input: Bipartite network of users, products, review ratings, helpfulness, verified purchased, review text.

Output: Score for every user (fraud), product (bad), review (fake)

Step 2: Grouping

findGroups()

Input: ranked list of users by score from Step 1, no. of top users k

Output: bot-users and products under attack

---

6. CURRENT PROGRESS

We have collected the review **data** from online sources such as Yelp and Amazon and have shaped it in the way our model accepts. Additionally, we have implemented **Loopy Belief Propagation** message passing algorithm for inference about the classification of nodes in the network. We have also tested the algorithm on small **test** review networks.

7. FUTURE WORK

Over the next few weeks, we will extend our current implementation by implementing pairwise Markov Random Fields for assigning probability to unobserved variables.

The nodes in the bipartite network will be linked by signed edges. The sign of the links will be inferred from attributes like rating, helpfulness, duplicates, and verified purchased tags.

HITS implementation for testing the improved FraudEagle framework (Signed Inference Algorithm).

We will compare our results with those obtained from previous SIA, and that of HITS.

References

[1] Mukherjee, A., Liu, B., Wang, J., Glance, N. and Jindal, N. 2011. Detecting group review spam. In *WWW.*

[2] Akoglu, L., Chandy, R. and Faloutsos, Christos. 2013. Opinion fraud detection in online reviews by network effects. In *ICWSM*

[3] Wang, G., Xie, S., Liu, B. and Yu, P. S. 2011. Review graph based online store review spammer detection. In *ICDM, 1242–1247.*