# Online Review Fake Detection Using FraudEagle with Prior Inclusions

## Abstract

Online Reviews on products (services) plays an influential role on the prospective customer decision on purchasing (availing) the product (service). This decides the fate of the product (services). Due to this there is need for an automated system to identify and curb such instances. However, it is a challenge for such machines to identify fake reviews. There are several techniques proposed and developed to identify such reviews and spammer groups. One such technique is the FraudEagle framework that exploits network effects among reviewers and products for fraud detection. In our work, we have extended this technique by adding new attributes to the review network such as *helpfulness*, *verified purchase* and *duplicates in reviews* for improved opinion spam detection which will discussed thoroughly in the further section.

Keywords: Opinion Spam, Fraud, Review

## Introduction

### Motivation

The problem of fraud detection is identifying fake reviews from a pool of user reviews posted online. This is a challenging problem as spammers post reviews that appear believable 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. Detecting such spam in online reviews is a difficult task. However there are many prior works which solved this problem using techniques like [5] exploited the review text duplicates in the reviews, [4] used burstiness in the reviews. The method proposed in [5] used duplicates in review as a positive training data to classify spam. The duplicates are detected using shingle method [6]. As this is a supervised model based on the training sets, it cannot be extended to other domains easily. For every other domain we need to collect data based on domain characteristics and train the classifier to identify the spam. Moreover selecting those specific sets is a tedious task, as datasets from different domains possess different characteristics. Exploiting burstiness in reviews specified in [4] is another approach to detect spam in reviews. In normal circumstances, the reviews arrive randomly. However there are certain situations where the number of reviews for the product are concentrated during a specific time period which is identified as review burst. The burstiness in reviews are detected using Kernel Density Estimation. However, on the downside of this approach is, it works only when there is a burst in reviews which is not commonly seen in online reviews.

And there are other novel approaches like [2], Identifying the spam based on the positive or negative sentiment in reviews. This algorithm uses review network undirected graph in which the users and products are nodes and edges are the relation between users and product based on the review, and it forms a bipartite graph with signed edges. Sign of the edge weight is based on the positive or negative sentiment in the review ratings. This algorithm is an unsupervised method which can be adapted to any domain. However, this method uses only the review rating feature to obtain the sentiment and correlations among users and products. There are other well defined features like usefulness count of a review, duplicates in review text, and verified purchase tags etc which will give accurate results, are not considered in the [2].

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

In the following sections, we have described the prior work, modified framework, evaluation and conclusion.

## 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. Candidate spamming groups are identified first followed by computation of spam indicator values. These groups are then ranked on the basis of indicators such as content similarity across the group, group size, group support, time window and whether the group posted early review in order to make a big impact. Similarly, Keystroke patterns have been used for deception detection by Banerjee et al. (Banerjee, Feng, Kang & Choi, 2014) by taking cues from editing manoeuvres and duration of pauses. Writing rate, pauses and revision rate have been measured from keystroke logs and used in detecting fake reviews. However, these approaches are largely based on and limited to the review text alone to detect the spamming behavior, which might not be working well when there is no review text available for a review.

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. On similar lines, 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. However, this algorithm have considered the review rating alone to predict the online spam.

## Our Contributions

In this work we extended the Fraud Eagle Framework proposed by Akoglu et al. (Akoglu, Chandy & Faloutsos, 2013), which is a network based technique to detect spam in online reviews. We used the below characteristics to compute the prior probabilities of user (product) to be honest (good) or fraud(bad) .

* Duplicates in reviews
* Usefulness Count of a review
* Verified Purchase Tag

These priors are used to compute the beliefs of each node (users/products) using Loopy Belief Propagation (LBP). LBP is an iterative message passing algorithm which works well for wide variety of applications. And to compute the probabilities we used pairwise Markov Random Field (pMRF) (Kindermann and Snell 1980) which works very well for an undirected bipartite graph.

### Duplicates Computation

In many online review we often tend to see the same review appearing multiple time, this one technique used by spammers to popularize/defame the product. So we tried to use this feature to compute the duplicate review count using cosine similarity. We considered the reviews are duplicated when the corresponding texts have a similarity score of 0.9. The correlation that can be drawn from this feature is, the more duplicate count the more likely the review to be spam and the reviewer to be fraud.

### Usefulness Count

The reviews from Amazon and Yelp has a field called usefulness which indicates that a particular review has been found useful by other potential customer who intends to purchase the product. This indicates that the review is review is genuine when the usefulness count is higher. Whenever there is no count available, we have given the default influence scores as 0.

### Verified Purchase

Amazon in its review provides a tag for every review called, verified purchase. This means that the user has purchased the product from their website. This is a very good indication of the user being genuine in writing the review and the corresponding review is more likely to be benign.

## Modified Framework

We used 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) is described with the three types of classes, Useful, Funny and Cool, each of this is associated with a count value indicating that some user(s) who wish to avail the service or purchase the product found this review useful or funny or cool. Out of these we decided to utilize only the useful count for this work, as Funny and Cool classes may not actually predict the review fairness. Duplicate reviews are detected using the cosine similarity, a shingle method. The duplicates are detected using a similarity score threshold > 0.9. Based on this, each user who wrote reviews for a product are given a score to indicate the untruthfulness of the 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 improved the existing framework by updating the priors based on the above characteristics.

**Outline of Modified SIA**

**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()

**Inpu**t: ranked list of users by score from Step 1, no. of top users k

**Output**: bot-users and products under attack

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.

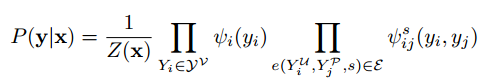
Labels, LU = {Honest, Dishonest} represents the domain of users and LP = {Good, Bad} represents the domain of products.

G(V, E), a signed review network graph in which users and products form the nodes. They are connected with signed links {+,-}.

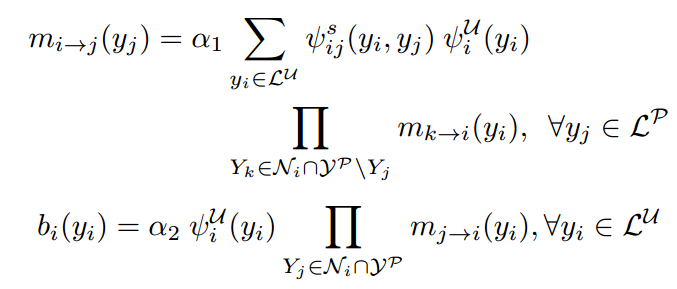
Let ΨiU Ԑ ѱ be a prior mapping ΨiU : LU  R ≥ 0,ΨjP : LP R ≥ 0 for each unobserved user Yi and unobserved product Yj

For each edge, let Ψijs Ԑ ѱ be a compatibility mapping.

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

****

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 Yi and product Yj 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.



Where mi → j is a message sent by user Yi to product Yj, α's are the normalization constants. LU denotes the label domain for users and LP denotes the label domain for products. bi (yi) of assigning Yi with label yi

## Dataset

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) of the reviews on the restaurants. We have also collected the academic dataset available from the website: <https://www.yelp.com/academic_dataset>

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

<< Graphs related to Data followed by the explanation of the graph >>

## Results

We have tested the modified SIA with a synthetic dataset as well as with the real data obtained from Yelp and Amazon.

<< To be updated based on the results >>

## Conclusion

In this project, we extended Signed Inference Algorithm (SIA) from [2], a graph based online spam detection algorithm by introducing more orthogonalities like helpfulness of the review, verified purchase tag etc. These orthogonalities are plugged in as priors to obtain the fake probability in reviews, products and users. We used a variant of HITS and native SIA to validate the results of this improved algorithm. Our experimental results using Yelp and Amazon datasets are promising. We have also validated using a supervised model generated using Logistic Regression. The Yelp Recommended and Non-Recommended reviews list to train the classifier as non-recommended reviews as positive training data. We tested this trained model with the top 50 users/reviews list obtained from our modified algorithm. And the results obtained are in good agreement with the supervised model. This modified algorithm has produced improved results over its predecessor and also over other similar techniques (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.*

[4] Geli Fei, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. Exploiting Burstiness in Reviews for Review Spammer Detection. In *ICWSM. 2013*.

[5] Jindal, Nitin, and Bing Liu. "Review spam detection."git Proceedings of the 16th international conference on World Wide Web. *ACM, 2007*.

[6] Broder, A. Z. On the resemblance and containment of documents. In Proceedings of Compression and Complexity of Sequences 1997, IEEE Computer Society, 1997.