# Visualization and Analysis of Restaurant Ratings Based on Neighborhoods

Zongrun Li
zli867@gatech.edu

Xuejie Guo
xguo324@gatech.edu

Shuheng Gan
shuheng.gan@gatech.edu

Siying Cen
scen9@gatech.edu

Jinyu Huang
jhuang472@gatech.edu

## 1 INTRODUCTION

There are many restaurant review sites and customers would like to find references from these websites and rely on online reviews to make decisions. People assume that these reviews are reliable. But reviews are sometimes fake. Yelp also admits 25% of reviews are fake[2]. Besides, people sometimes just have general ideas when they find restaurants such as I would like to eat American food. Neighborhoods which have higher density and ratings of restaurants will help people find fit restaurants. However, none of restaurants provide such information to customers.

Visualization and analysis of restaurant ratings based on neighborhoods is a project which dedicates to provide users with fair ratings for restaurants in Atlanta region by collecting and analyzing data from Yelp and using visualization methods based on neighborhoods. Users can find accurate and easy-understanding information such as rating distribution, density of restaurants from our interactive figures. They can also choose restaurant category, rating range and price range to find restaurants.

## 2 PROBLEM DEFINITION

Neighborhood is a common category that people used to decide where to eat and each neighborhood may have some unique demographical features that may influence the consumption of the food. Meanwhile, fake reviews exists and there are few restaurant website provide users with neighborhood information. The problem can be defined as exploring how restaurants' features will be influenced by the neighborhood which it is belonged to and building the platform that customers can make their choices based on neighborhoods' information. And we also explore whether a website which provides users with neighborhoods information is better or not. To solve this problem, we need firstly classify restaurants into neighborhoods and identify the fake ratings. Then analyzing the relationship between restaurants' rating and neighborhood. Finally, design a interactive website based on neighborhoods' data and analyzed information.

## 3 SURVEY

Ratings and visualization methods affect customers' behaviors. For ratings, Luca proved the causality effect of Yelp ratings on the revenue of restaurants with regression approach. Additionally, the impact of ratings is correlated with many other features such as number of reviews and certified reviewers[12]. To provide fair ratings, Jiang combined NLP with K-medoids clustering to cluster Yelp's review which may make us find different reviews' features[7]. Jindal and Liu detected duplicate reviews by using Jindal's detection model[9]. Xie et al., Lim et al. and Mahmudur et al. found that fake reviews used to be generated in a very short period of time[11, 15, 20]. Mahmudur established Marco system to catch fake review in Yelp which based on RSD, ARD and FRI modules[15]. And Mahmudur et al. also used another method based on CoReG, RF, IRR and JH algorithms to detect fraudulent behaviors in Google Play[19]. After recognizing the review spikes, Andrew et al. provided a mathematics method to detect abnormal ratings based on Bayesian reputation systems [16]. Li et al. and Oh et al. designed algorithms based on reviewers' attributes[6, 14]. Li et al. considered reviewers' IP and features [6]while Oh et al. adjusted the reputation based on the confidence of customer ratings which has been calculated based on customers' activity, objectivity and consensus[14]. Masound's compared different recommendation algorithms and provided us insights about how to choose fair recommendation algorithms[13]. Kim et al. dedicate to find a clear criteria for evaluating users in Yelp by using logistic

regression, PageRank and SVM algorithms[10].

All the algorithms for detecting fake reviews have shortcomings because fake reviewers will change their methods to avoid their reviews be deleted based on fake detection algorithms. Some methods dedicate to find unreliable reviewers and some methods dedicate to find review spikes to avoid fraudulent reviewers. We decide to combine these two ideas to provide fair ratings.

For visualization methods, Yang et al. found that presenting information in a hierarchical way makes it easier to capture information[21]. And Jin et al. proved that labeling restaurants with keywords of reviews will help our users know the general comments and make decision efficiently[8]. Zhang et al. used DBSCAN to identify the center of their check-in area and then analyze the overall check-in probability over spatial distance. They found out that most users' check-in travel is no more than 2 km[1]. Sun et al. created Voronoi diagram to cut the area into regions of venues, then run A Multidirectional Optimum Ecotope-Based Algorithm to analyze the influence of geographical features to venue's rating and revealed that the spatial features will influence the rating of certain kind venues[17].

These researches use mathematics methods to classify and divide areas. But actually, people are unfamiliar with these newly defined area. The natural way to divide cities are neighborhoods.Since that, we will consider about visualizing restaurants' distribution and ratings information based on neighborhoods to help customers make choices easily and to avoid the shortcomings.

# 4 PROPOSED METHOD

## 4.1 Intuition

*4.1.1 Innovations.*

1. Restaurants are classified based on neighborhoods. It is a friendly way for costumers who only has general ideas to find a place to eat for a specific category of restaurants.

2. The restaurants' ratings are not simply based on all ratings' average. We use algorithm to detect days which have fraudulent behaviours and fake reviews in these days.

3. The project visualizes results from our analyses. Restaurants with different ratings and price have markers in different colors and sizes. Customers can also filter restaurants easily based on interactive website.

## 4.2 Approaches

*4.2.1 Yelp Data Collection.*

Yelp Fusion API and Google API are useful tools for getting restaurants' and users' data. However, Yelp API can only return up to 1000 results per query and only returns three selected reviews for each restaurant. Same as the Yelp API, Google API only returns 20 data per query. So, we wrote a program to scrape the users' features from Yelp web page to and used the Yelp API to get restaurants' data.

We separate data extraction into two main steps: web crawl and web analysis. We first send the http request to get server response, then pre-process the scrapped HTML file with Beautiful soap. Then extract the users' features by analysing the website structure.To prevent being block, we set timer and use random user agent.

We collected 2356 restaurants' data by using Yelp Fusion API and 71432 reviews data by web scraping.

*4.2.2 Data processing.*

The original data scratched from Yelp needs cleaning before being processed. For example, we observe duplicated notations of restaurant type, which is expected to be a significant predictor of ratings. In order to merge the similar notations, we firstly intend to apply natural language processing method to forming a distance matrix, based on which to carry out clustering. In practice, we observe that the clustering results don't meet our expectations, as the within-cluster restaurant types have little in common with respect to real-life meaning. We then choose to determine restaurant types manually by referring to their linguistic meanings, geographical origins and characteristics. For example, French restaurants and German restaurants can be clustered into one type named European restaurants. By doing so, the total number of restaurant types is reduced from 147 to 20.

What's more, we need to merge the review dataset and restaurant dataset together in order to ease the reference between one and another. For this part, we make use of the Pandas module in Python for implementation.

*4.2.3 PNPoly.*

To efficiently offer users ideal restaurant choices and improve the entire dining experience, we decided to group restaurants into smaller and more specific

zones. We decided to use neighborhood unit because the planning and economic development between neighborhoods are quite different[4]. After we collected the restaurant coordinates and neighborhood borders data, we transferred this to a classical point-in-polygon (PIP) problem, i.e. how to decide a given point is inside or outside of a polygon. Crossing number algorithm and winding number algorithm are two common solutions to PIP problem[3, 5, 18]. Since crossing number algorithm is very straightforward and the other one is preferred for a nonsimple closed polygon (e.g. one that overlaps with itself, which is not suitable for our 2D map situation), we chose the former method. The logic of crossing number algorithm is: if a point is inside of a polygon, any ray from this point will have odd intersection(s) with edges of the polygon; if outside, even (include 0) intersections. The Franklin's point inclusion in polygon (PNPOLY) algorithm[3] will test each edge of a polygon and count the intersection number when there is at least one. To avoiding redundancy, it always chooses a horizontal ray parallel to the positive x axis and pointing to the right of a point. Thus, when a point is on the left boundaries, it is considered to be inside; on the contrary, it is outside.

#### 4.2.4 *Fake Review Detection.*

To detect fraudulent behaviours, we assume that fraudulent behaviours will continue in a relative short time(one day) and fake reviews happen on these days. Since that, we will first detect days with higher numbers of positive reviews than normal and then detect the fake reviews in these days. To detect abnormal days, Mahmudur Rahman1 et al. indicate that when the reviews number greater than 1/7 of all reviews number before that day, the ratings will obviously increase[15]. So, we will first select abnormal days using that condition. For fake reviewers detection, we assume there are enough fake reviewers on abnormal days and we also assume elite members are reliable. We use this assumption to label reviewers and use reviewers' attributes which have related to reviewers' credibility such as friends, photos, etc.[10]. We use Random Forest to train the model and use the model to judge whether a reviewer reliable or not. After judgement, we will delete unreliable reviewers' ratings on abnormal days and calculate the ratings as our result.

#### 4.2.5 *Feature Deduction.*

We will use random forest regression algorithm to evaluate correlations between restaurants features and ratings. To predict restaurant rating, features such as location, neighborhood, total review numbers, opening hours and category will be extracted from dataset; linear regression algorithm will be performed to create rating prediction model. Training and testing dataset are separated from the whole dataset. Once the model is constructed, we can figure out most relevant features to rating. This part can help us know what features affect customers' ratings. Based on the result, customers especially merchants can figure out where is a good place for restaurants.

#### 4.2.6 *Visualization.*

The website has two pages. Main page which includes a main-plot and a subplot and an About page which includes some information and user manual for the project.

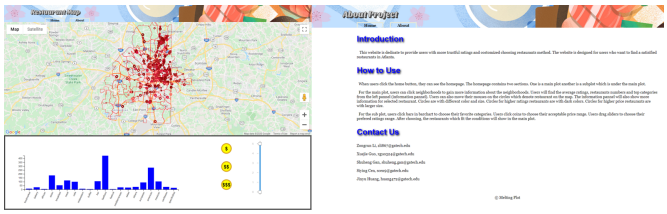


**Figure 1: Website**

For the main page, the main plot is based on google map API and JavaScript. The Atlanta map are divided by different neighborhoods. Users can select neighborhoods to get more information including restaurants numbers, averaged ratings and top category. The information displays on the right side of website (information panel).

Restaurants with different ratings and price have different markers with different colors and sizes. Users can select these markers to have more information about restaurants from information panel.

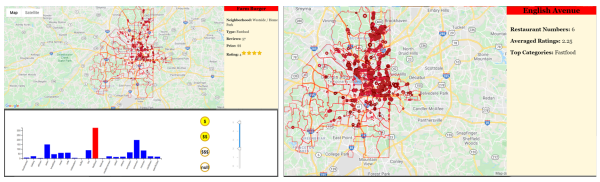

**Figure 2: Selecting Features**

The subplot is based on d3.js. The subplot provide users with restaurants' categories and price selector. Users can filter restaurants by using subplot. For any users' operations, the main plot will change and the bar chart will also update based on users' choices. Users can see the density of filtered restaurants easily and it is convenience for them to choose a neighborhood to find restaurants which fulfill their requirements.

# 5 EXPERIMENTS/ EVALUATION

Our project includes four experiments, PNPloy, fake review deduction, feature detection and data visualization.

## 5.1 PNPoly

We use PNPoly algorithm to classify each restaurant into a neighborhood unit based on their coordinates. As most restaurant are located in midtown or downtown area, we combine several adjacent neighborhoods that lack for restaurant data together as one unit. The total number of neighborhoods in Atlanta, Georgia is 242 officially, while in our experiment we use 102 neighborhood units. For each neighborhood unit, we treat it as a polygon and used a three dimensional array to store the latitude and longitude of each vertex. And we loop all neighborhood units for each restaurant to label it a 'classified_neighborhood' attribute (may different from the original neighborhood attribute that Yelp offers). We also make sure that every restaurant only be classified into one unit. By comparing our results with Yelp raw data, we find that the neighborhood attribute in Yelp is not directly related to the restaurant location. Some restaurants are labelled with multiple neighborhoods, especially those with very few reviews. This method will increase the probability of a restaurant being visited. However, it will produce redundant information and may cause inconvenience to users who have clear dining destination. We throw away 640 restaurant data that don't belong to any Atlanta neighborhood unit and finally kept 1,716 restaurants. In the same time, we calculate the restaurant number, average restaurant rating, and the top category for each neighborhood unit. The original data from Yelp contain 213 restaurant categories. Actually some are the subcategories of others. We organize them manually and finally get 20 categories. In nearly half neighborhoods (n=40), the top category (the category of most restaurants) is "fast food". The neighborhood unit "Castleberry Hill, Downtown" has the largest number of (n=319) restaurants and "Midtown" has the second most (n=201). The rest neighborhood units have less than 100 restaurants.

## 5.2 Fake Review Detection

We will answer which days have fraudulent behaviours and who are reliable reviewers on these days. Based on each days reviews number for each restaurants, we will use spike detection method to find abnormal days as we mentioned before and we find 1235 spike days by using our dataset. And we assume fraudulent behaviours happened only on these days.

To make these days ratings reliable, we will use Random Forest to find who are reliable reviewers. For Random Forest, we will label elite users as reliable users and the others are unreliable. The dataset is from abnormal days and attributes are from data we scratched such as reviewers' friends number, reviews number, photos, etc.. We will use 80% of data to train the model and 20% to test the model.

The ratings dataset covers information of 47502 unique users, who make ratings for 97 restaurants. The 'Elite' variable is the label, which equals 1 for elite users and equals 0 for non-elite users. We use features including 'friends', 'number_reviews', 'photos' and 'location' in order to make prediction. For the 'location' variable, users who is located in GA get the value of 1 and those who are not get the value of 0. Making prediction based on the test data, we obtain accuracy of 93.6%. After that, we fit the whole dataset and get a series of predicted labels. As mentioned before, we assume the ratings from elite users are reliable. In that case, we filter out the elite users using the predicted label and calculate the mean ratings for each restaurant.

In comparison of the original ratings (taking account of all users), original elite ratings (determined by the elite users given by raw dataset) and predicted elite ratings (determined by predicted elite users), we find that the predicted elite ratings tend to be higher than the other 2 in general by referring to the mean. It is also worth pointing out that the signs of differences between predicted ratings and original ratings are not consistent. That means fake ratings can be either higher or lower than the fair rating. For example, the restaurant with id

'iNpKyGdjCb40PECLpyF-rg' has the highest value of rating difference, which is 2.19 (4.69 for the observed original rating and 2.5 for the predicted rating). Based on this, we are suspicious of the original rating and regard the predicted rating as more reliable.

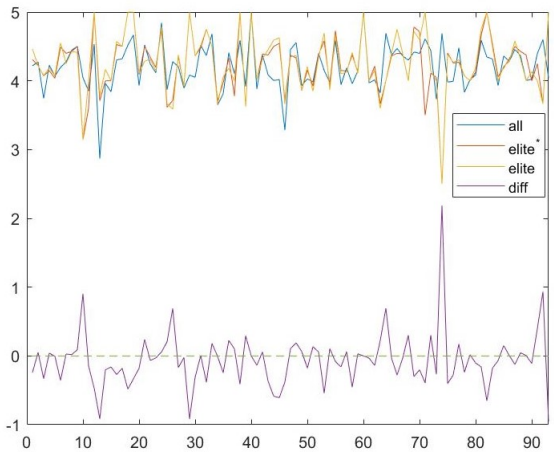


**Figure 3: Random Forest Result**

## 5.3 Feature Deduction

For the rating analysis, we investigate the relation between yelp user information and their rating deviation of a specific restaurant. Five features were selected to define a yelp user (num of friends, num of reviews, num of photos, living area, elite user or not) and fit with their rating deviation to the rating mean of one restaurant. Data were collected from 71432 yelp users who rated 97 restaurants in Atlanta in total.

$$user\ rating\ deviation = \frac{rating - mean(rating)}{mean(rating)}$$

Binary vector were used for representing the user area (in Atlanta or not) and 'elite2020' label (elite user or not).

The results of random forest regression showed that before fake date detection, Mean Absolute Error equals to 0.22143776405341808

We found that the feature (number of friends) influences the rating deviation most, while label "elite" influences the least. So we are suspicious of the reliability of Yelp's "elite" label which validate that we need to find a new criteria for elite members just as we do in fake reviews detection.

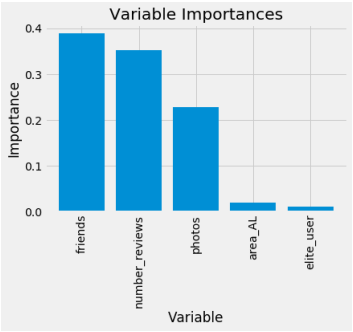


**Figure 4: Bar charts of feature importance of user information in predicting user rating deviation after fake date detection (random forest n_estimator = 20).**

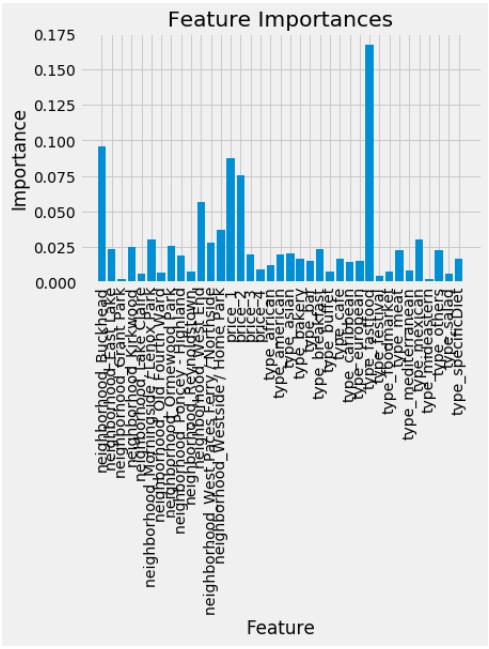


**Figure 5: Bar charts of feature importance of restaurants in predicting rating for recommendation (random forest n_estimator = 100).**

For restaurant recommendations, we performed random forest regression to predict restaurants' average rating with restaurant features (neighborhood, price level, restaurant type) from 2351 restaurants in Atlanta.The features with the highest importance factor are: type_fastfood (0.1677), neighborhood_Buckhead (0.0957), price_$ (0.0875), price_$$ (0.0754), neighborhood_West End (0.0563) and neighborhood_Westside / Home Park (0.0369).

We found that attributes of 'fastfood́type, 'price_$' and 'Buckhead' neighborhood had the highest correlation to restaurant rating. It is obvious that eating fastfood or eating at these neighborhoods are good choice for customers. And for merchants, it is wise to have a fastfood restaurant in these neighborhoods.

## 5.4 Data Visualization

We will answer which neighborhoods are best for finding a restaurants under customized conditions by using our data visualization website. For example, we want to find a good neighborhood for higher ratings and lower price Asian food. We select on the website based on requirements and we can easily see midtown is a good neighborhood to go. Additionally, it is easy to figure out which neighborhoods have higher restaurants density. And for costumers, it is convenience to know which restaurants have higher ratings or higher price since markers' colors are related to ratings and markers' size are related to price (Darker color means higher ratings and smaller size means lower price while null price is with the smallest size). From bar chart, we can easily see how different types of restaurants' numbers varies during we change the price and ratings range.

The visualization website also validate that restaurants are concentrated in some neighborhoods, such as midtown. And fast food restaurants obviously have the highest number. It may be because fast food is popular as we analyzed before.

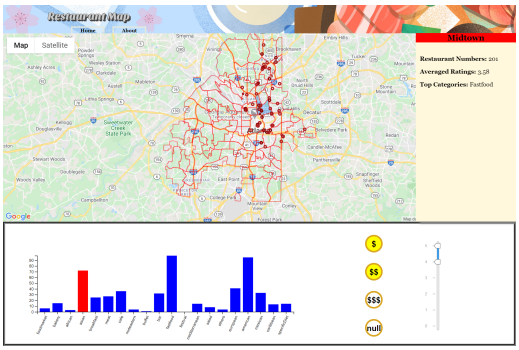


**Figure 6: Neighborhoods for Asian Food**

# 6 CONCLUSION AND DISCUSSION

## 6.1 Effort

All team members will contribute similar amount of effort.

Xuejie is mainly responsible for data collection. Shuheng is mainly responsible for PNPoly algorithm. Jinyu is mainly responsible for data processing and fake review detection. Siying is mainly responsible for feature deduction. Zongrun is mainly responsible for visualization. Thanks for all teammates' contributions.

## 6.2 Conclusion and Discussion

In PNPoly analysis, we realize the neighborhood attribute in Yelp does not necessarily show the real restaurant location. Some new restaurants or ones with few reviews may be labeled with multiple neighborhoods to raise the chance of being searched and therefore being visited. However, this strategy may increase useless information for users. Thus, we relabeled the neighborhood attribute for each restaurant based on PNPoly algorithm. Our results showed downtown and midtown are the most popular neighborhoods and fast food is the most common type of restaurant in Atlanta.

For the fake review detection part, our focus is to provide a more reliable rating for a specific restaurant rather than explicitly tell which review/reviewer is fake. We use random forest to give our own classification of elite users and non-elite users, and filter out the reviews made by elite users only, based on which to recalculate ratings. As we can observe from the experiment results, some abnormal ratings can be revealed through this approach. That is to say, the reviews/ratings made by elite users on Yelp can be regarded as a relatively reliable reference. With future endeavor, we will collect more data, make use of more advanced methodologies like natural language processing, count on more features to identify fake reviews in a more explicit way.

In feature deduction, important features for user review deviation and restaurant rating were identified, providing reference for detecting reliability of user review and restaurant recommendation based on neighborhood, food type and price. In Atlanta, a fast food restaurant with cheaper price and located in Buckhead neighborhoods is more likely having a higher rating.

For visualization part, we know restaurants are concentrated in some neighborhoods, especially in downtow. And in Atlanta, fast food resturants number is the highest. Users by using this website can easily find neighborhoods and restaurants which fuilfill their requirements.

# REFERENCES


[1] 2016. *WWW '16: Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE.

[2] BBC. [n. d.]. Yelp admits a quarter of submitted reviews could be fake. www.bbc.co.uk/news/technology-24299742.

[3] W Randolph Franklin. 2006. Pnpoly-point inclusion in polygon test. http://www.ecse.rpi.edu/Homepages/wrf/Research/Short_Notes/pnpoly.html. Accessed: 2020-03-31.

[4] Atlanta Government. [n. d.]. Neighborhood Gentrification Pressure Areas. https://www.atlantaga.gov/home/showdocument?id=33833. Accessed: 2020-03-28.

[5] Eric Haines. 1994. Point in polygon strategies. *Graphics gems IV* 994 (1994), 24–26.

[6] Bing Liu Xiaokai Wei Huayi Li, Zhiyuan Chen and Jidong Shao. 2016. Spotting fake reviews via collective positive-unlabeled learning.. In *2014 IEEE international conference on data mining*.

[7] Yimin Liu Jiang, Renfeng and Ke Xu. 2015. A general framework for text semantic analysis and clustering on Yelp reviews.

[8] Wei Jin, Hung Hay Ho, and Rohini K Srihari. 2009. OpinionMiner: a novel machine learning system for web opinion mining and extraction. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1195–1204.

[9] Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *Proceedings of the 2008 international conference on web search and data mining*. 219–230.

[10] Caroline Kim, Gordon Lin, and Honam Bang. 2015. Discovering Yelp Elites: Reifying Yelp Elite Selection Criterion. *University of California-San Diego* (2015).

[11] Ee-Peng Lim, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady Wirawan Lauw. 2010. Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. 939–948.

[12] Michael Luca. 2016. Reviews, reputation, and revenue: The case of Yelp. com.. In *Harvard Business School NOM Unit Working Paper*. 12–16.

[13] Bamshad Mobasher Robin Burke Mansoury, Masoud and Mykola Pechenizkiy. 2019. Bias disparity in collaborative recommendation: Algorithmic evaluation and comparison.. In *arXiv:1908.00831*.

[14] H. Oh, S. Kim, S. Park, and M. Zhou. 2015. Can You Trust Online Ratings? A Mutual Reinforcement Model for Trustworthy Online Rating Systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 45, 12 (Dec 2015), 1564–1576. https://doi.org/10.1109/TSMC.2015.2416126

[15] Mahmudur Rahman, Bogdan Carbunar, Jaime Ballesteros, and Duen Horng (Polo) Chau. 2015. To catch a fake: Curbing deceptive Yelp ratings and venues. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 8, 3 (2015), 147–161. https://doi.org/10.1002/sam.11264 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/sam.11264

[16] Mahmudur Rahman, Mizanur Rahman, Bogdan Carbunar, and Duen Horng Chau. [n. d.]. *FairPlay: Fraud and Malware Detection in Google Play*. 99–107. https://doi.org/10.1137/1.9781611974348.12 arXiv:https://epubs.siam.org/doi/pdf/10.1137/1.9781611974348.12

[17] Jorge David Gonzalez Sun Yeran Paule. 2017. Spatial analysis of users-generated ratings of yelp venues. *Open Geospatial Data, Software and Standards* 2 (2017). https://doi.org/10.1186/s40965-017-0020-9

[18] Dan Sunday. 2012. Inclusion of a Point in a Polygon. http://geomalgorithms.com/a03-_inclusion.html. Accessed: 2020-04-01.

[19] Andrew Whitby, Audun Jøsang, and Jadwiga Indulska. 2005. Filtering Out Unfair Ratings in Bayesian Reputation Systems.

[20] Sihong Xie, Guan Wang, Shuyang Lin, and Philip S Yu. 2012. Review spam detection via temporal pattern discovery. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 823–831.

[21] S. J. Woo Y. J. Ah, K. H. Jeong and L. S. Ho. 2011. Visualization of restaurant information on web maps.. In *The 5th International Conference on New Trends in Information Science and Service Science, Macao*.