A Tangled Web: Evaluating the Impact of Displaying Fraudulent Reviews

Completed Research Paper

Uttara M. Ananthakrishnan
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
uttara@cmu.edu

Beibei Li
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
beibeili@andrew.cmu.edu

Michael D. Smith
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
mds@cmu.edu

Abstract

The growing interest in social media for legitimate promotion has been accompanied by an increasing number of fraudulent reviews. Beyond fraud detection, little is known about what review portals should do with fraudulent reviews after detecting them. In this paper, we study how consumers respond to potentially fraudulent reviews and how review portals can leverage such knowledge to design better fraud management policies. To do so, we combine randomized experiments with statistical learning using large-scale archival data from Yelp. Our experiments show that consumers tend to expand the variety of their choice set during product search and to increase their trust towards the review portal when it displays fraudulent reviews along with non-fraudulent reviews, rather than censor fraudulent information. Finally, our archival analysis using a Maximum Likelihood Estimation method allows us to design a novel fraud-awareness reputation system that platforms can deploy to better improve consumer trust and decision making.

Introduction

"Oh what a tangled web we weave,
When first we practice to deceive"
- Sir Walter Scott, 1808, Marmion: A Tale of Flodden Field,

With over half of the Internet's population accessing social media sites regularly, online reviews have become a major source of information for consumers. Prior work has found significant economic effects from online reviews in variety of settings, including the effect of reviews on product sales and purchase decisions (e.g., Mayzlin et al. 2012), the effect of rating on restaurant bookings (e.g., Anderson and Magruder 2012) and hotel bookings (e.g., Ghose et al. 2012). Organizations are searching for new ways to engage more followers via social media and to garner more positive reviews than their competitors.

However, this increased interest in social media for legitimate promotion has been accompanied by an increasing number of fraudulent reviews on major online product search websites, such as Amazon (e.g., Jindal et al. 2010), TripAdvisor (e.g., Mayzlin et al. 2012; Ott et al. 2011), and Yelp (e.g., Luca and Zervas 2013) as some companies choose to pay for favorable reviews to create an illusion of consumer loyalty and customer advocacy on social media sites. A recent report by research firm Gartner estimated that by 2014, 10% to 15% of all social media reviews are fraudulent (Gartner 2013).

To improve the credibility of the online reviews, product search engines have started combining techniques from text mining and natural language processing (e.g., Jindal and Liu 2008; Ott et al. 2011) as well as network-based graphical models (e.g., Akoglu et al. 2013) to detect potentially fraudulent reviews on their websites. However, beyond this literature on detecting fraudulent reviews, little is known about what review portals and search engines should do with the fake reviews after detecting them. This is the subject of our research.

Currently, most review portals choose to deal with fake reviews by quietly deleting them. However, certain review portals like Yelp, go one step further: invalidate suspected fraudulent reviews and make the review visible to the public with a notation that it is potentially fraudulent. There is, of course, a natural tension between the two approaches: Displaying fraudulent reviews aims to discourage fraud by reducing information asymmetry and increasing the reputation risks of being detected. The hope is that the expected cost of being caught cheating will outweigh the potential revenue increase associated with a small rating increase. However, this approach can be risky for social media platforms because it may highlight the prevalence of fraud, and affect consumers' trust in all reviews.

We address this managerial challenge by answering the following two questions in our research:

1. What is the impact of suspected fake reviews on consumer behavior on a website?

2. Is there a more effective fraud management policy for social media and search engine platforms when responding to potentially fake reviews, i.e., to discourage sellers from soliciting fraudulent reviews?

First, we are interested in the impact of the fraudulent reviews. In particular, we want to explore whether displaying fraudulent reviews can impact consumers' behavior, and how they may influence consumers' trust on a review portal. Fraudulent reviews can be viewed as similar to advertisements or persuasion by businesses (or their competitors), except that the senders' true identities and incentives are uncertain (Mayzlin et al. 2012). Advertising literature also suggests that high-quality sellers are likely to signal their quality through intensive advertising only when advertising costs are significantly high (e.g., Milgrom and Roberts 1986; Nelson 1974). In the case of posting fake reviews, the cost is relatively small. As a result, the seller loses little by claiming his product quality is high. Given the expected high benefit, a low-quality seller may be more motivated to do so compared to a seller who has already established higher quality reputation (e.g., Luca and Zervas 2013; Mayzlin et al. 2012). Therefore, a larger portion of fake promotional reviews may in fact indicate lower quality of the seller, which in turn may engender mistrust among consumers towards the ethics of the seller leading to a lower demand for its products. This theory also operates under the assumption that consumers can efficiently discern fraudulent reviews from non-fraudulent reviews. However, Ott et al. 2011 discuss how consumers find it very hard to differentiate between these two kind of reviews just based on textual information.

As noted above, theory alone does not conclusively indicate how the presence of fraudulent reviews may convey the quality signal, and more importantly how it may influence product demand through its impact on consumer beliefs about the product quality. In our paper, we aim to examine these questions from empirical and experimental perspectives. We are interested in examining questions such as: How does the exposure to fake reviews in parallel with the truthful reviews affect consumer behavior and choice on a website? How does the composition of the fake reviews (positive vs. negative) indicate the quality of the review portal? Moreover, beyond the direct economic impact on the product demand, it is not clear how fake reviews may affect the potential trust of consumers towards the review portal, which may have indirect economic impact on both product demand and the review portal in the long run. In our paper, we study this effect in the context of fake reviews and whether consumers perceive a portal as more trustworthy when the fake reviews are displayed instead of being deleted.

Finally, we aim to understand what a review portal should do once it identifies a fraudulent review. We are interested in answering the following managerial policy questions: *Should a review portal inform its users about the potential fraud reviews on its platform? Should the review portal display the suspected fraud reviews on its platform after detecting them or should it silently delete them? How can review portals discourage fraudulent reviews and better convey product quality by taking into account sellers' dishonest behavior?*

To achieve our goals, we combine two randomized user experiments based on a restaurant review portal we designed and implemented ourselves, together with archival data analysis using a large-scale dataset we collected from Yelp.com containing 283,830 fraudulent and non-fraudulent reviews for 982 restaurants in San Francisco. Our goals and main findings can be summarized as follows:

1. The goal of our first randomized experiment is to understand if there is a significant change in user behavior when fraudulent reviews are displayed along with the non-fraudulent reviews.

    - We find that when fraudulent reviews are displayed along with the non-fraudulent reviews users do behave significantly differently in a manner suggestive of higher user engagement. Under this condition, users are more likely to expand their choice set during search by visiting more restaurants, and spend more time on the review portal.
    - Moreover, we find that consumers are more likely to choose restaurants that have a lower historical fraudulent activity when fraud information is displayed to them versus when this information is not displayed. This finding suggests that displaying suspected fraudulent reviews information could potentially help consumers improve their decision making.

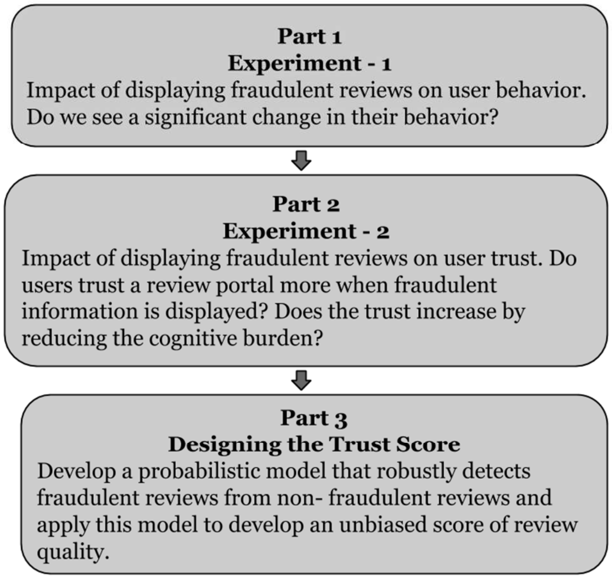
2. The primary aim of the second experiment is to understand the impact of fraudulent reviews on the trust that a user places on the review portal. Using techniques from behavioral economics, we conduct a second randomized experiment to quantify the user's trust for website using a betting game.

    - From this experiment, we find that displaying fraudulent reviews and providing a summary score of potentially fraudulent activities causes users to trust a review portal more than they otherwise would.

3. Finally, based on the findings from the two experiments, we propose a design for a novel fraud-awareness reputation system. In particular, we collected a dataset from Yelp.com containing fraudulent and non-fraudulent reviews over two years among restaurants in San Francisco. We propose a Maximum Likelihood Estimation model to estimate the fraud probability of each review conditional on the characteristics of textual reviews, reviewers and restaurants. We observe that the accuracy of our model increases drastically by adding features pertaining to the reviewers and the restaurants when compared to just using textual information. This strengthens our belief that users are not likely to be very efficient in identifying fraudulent reviews by just reading the textual information. It highlights the importance of having a robust fraud detection model that can combine reviewer and restaurant level features for detecting fraud. Estimates from the final model allow us to derive a trust score for each product by enforcing a penalty on the detected fraud reviews. The basic idea behind our proposed trust score is to incorporate a reputational cost for every fake review. Mayzlin et al. 2012 and DellaVigna and La Ferrara 2010 argue how the rate of manipulation is affected by the reputation cost of the players involved. Unfortunately, the existing approaches by

review portals to dealing with fraud do not include such a penalty. Our model provides an innovative way in which review portals can learn from the fraudulent reviews and better manage them to increase user engagement and consumer trust.

A summary of our paper's structure is provided in Figure 1.



Part 1
Experiment - 1
Impact of displaying fraudulent reviews on user behavior. Do we see a significant change in their behavior?
Part 2
Experiment - 2
Impact of displaying fraudulent reviews on user trust. Do users trust a review portal more when fraudulent information is displayed? Does the trust increase by reducing the cognitive burden?
Part 3
Designing the Trust Score
Develop a probabilistic model that robustly detects fraudulent reviews from non- fraudulent reviews and apply this model to develop an unbiased score of review quality.

Figure 1. Summary of Analysis

Literature Review

Our online fraud review question relates to four different streams of the literature, spanning multiple domains and dealing with different aspects of the problem.

The first set of papers establish the importance of online reviews and how information sharing about an online shopping experience helps increase trust between customers and merchants (Ba and Pavlou 2002; Chevalier and Mayzlin 2006; Dellarocas 2003; Resnick and Zeckhauser 2002). The second set of papers look at the impact of online reviews on consumer behavior, product evaluation and purchase decision (e.g., Ahluwalia and Shiv 1997; Chatterjee 2001; Chen and Xie 2008; Kanouse 1984; Lascu and Zinkhan 1999). Novel techniques have been developed to mine and predict the usefulness and the subjectivity of these reviews based on various parameters which provides context to what reviews have the higher probability of influencing a customer's purchase decision (e.g., Ghose et al. 2012; Netzer et al. 2012; Yatani et al. 2011 and Archak et al. 2011). However, existing literature does not address fraudulent reviews or their impact in consumer's decision making. In our paper, we focus on the impact of fraudulent reviews on a consumer's behavior on a review portal and identify best strategies for organizations in managing fraudulent reviews.

The second set of papers is predominantly from computer science literature, which has an extensive collection of work done in the fraud detection domain (e.g., Jindal and Liu 2008; Lim et al. 2010; Mukherjee and Liu 2012; Wu et al. 2010; Yoo and Gretzel 2011). Most of them involve looking at user and/or review characteristics, anomalies in posting patterns (Jindal and Liu 2008; Jindal et al. 2010) and some use scoring methods (Lim et al. 2010) to look for review spam. In this paper, we propose a probabilistic model that robustly identifies fraudulent reviews. We use this model in proposing a trust score based system that will help the review portals in conveying their ability to detect fraud and also penalize businesses that resort to dishonest practices. There is also some work in the Human Computer Interaction(HCI) literature on measuring consumers' trust while adopting a website (e.g., Jensen et al. 2000; Riegelsberger et al. 2003; Zheng et al. 2002). This body of literature looks at the question of how to identify fraudulent reviews without going to details about their economic or managerial implications. Firms are increasingly using sophisticated algorithms from this body of research, especially using Machine Learning and Natural Language Processing to identify fraudulent reviews. While this body of literature plays a very important role in the life of a fraudulent review, it does not address what should be done once these fraudulent reviews are identified, which is the main focus of our research.

The third set of papers is particularly important to our problem. These papers look at the economic aspects of review fraud manipulation that happens on review platforms like Yelp, Expedia and Trip Advisor. For example, the work by Mayzlin et al. 2012 analyzes the impact of a verification mechanism on the cost of leaving a fake review in Expedia versus Trip Advisor and the motivation to commit fraud. Luca and Zervas 2013 investigate the motivation for a restaurant to solicit fake reviews. However, these two papers look mainly at the antecedents and motivation for businesses to commit fraud. This body of research does not look at fraudulent reviews from the consumer's perspective, nor provide suggestions on how to manage fraudulent reviews once the motivation is identified.

Therefore, our research question fits well in the later stage of a fraudulent review's life cycle (Stage 3 and Stage 4 marked in red in Figure 2) which none of the existing papers have looked at. After identifying fraudulent reviews, the responsibility of determining the fate of the fraudulent review lies solely in the hands of the review portal. The natural tension between displaying these fraudulent reviews to its consumers to earn their trust and not displaying these fraudulent reviews and risking more future fraudulent behavior raises an important managerial question which, to the best of our knowledge, none of the existing papers have addressed. Our paper is the first to tie the impact of informing users about review spam on consumer behavior and trust on the website. We are the first to look into efficient fraud management techniques that the websites could apply once fraud reviews are identified not only to increase the consumer trust but also to reduce the cognitive burden in processing fraud information. Existing systems for fraud management do not have a reputational cost associated with detection. Prior work (e.g., Mayzlin et al. 2012 and DellaVigna and La Ferrara 2010) argues that such a system is very effective in discouraging future dishonest behavior. In this paper, based on the rich dataset, we finally propose a technique to design a score-based metric that would not only help the website to signal its effectiveness in keeping fraud reviews at bay, but also enforces a penalty for businesses that resort to soliciting fake reviews.

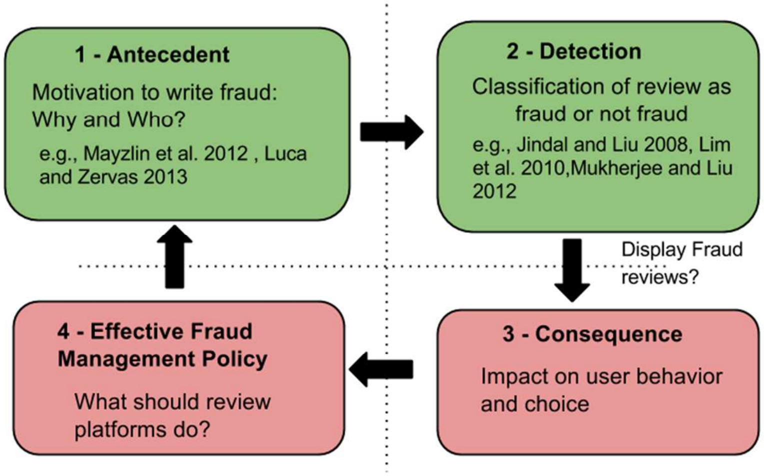

1 - Antecedent
Motivation to write fraud: Why and Who?
e.g., Mayzlin et al. 2012 , Luca and Zervas 2013
2 - Detection
Classification of review as fraud or not fraud
e.g., Jindal and Liu 2008, Lim et al. 2010,Mukherjee and Liu 2012
Display Fraud reviews?
3 - Consequence
Impact on user behavior and choice
4 - Effective Fraud Management Policy
What should review platforms do?

Figure 2. Life Cycle of a Fraudulent Review

Experiment I - Impact of Displaying Fraudulent Reviews on User Behavior

The majority of review portals choose not to display suspected fraudulent reviews once they are detected. These suspected fraudulent reviews are simply removed from the consumers' purview. One of the main reasons to remove fraudulent reviews might be that consumers' decision making and behavior on a review portal is not going to be influenced by fraudulent reviews and therefore showing them the non-fraudulent reviews might be enough. We wanted to understand if this hypothesis is true, i.e., do consumers remain unaffected by the presence of fraudulent reviews or can we observe any change in their behavior and choice by displaying fraudulent reviews. To achieve this, we designed and conducted a randomized experiment on Amazon Mechanical Turk between February 21, 2014 and March 4, 2014 with 554 subjects.

In particular, we wanted to see how displaying fraudulent reviews in review portals performed against different baseline approaches including the currently popular "Silent Approach", in which the portal does not mention the measures it takes against fraudulent reviews or even acknowledge the possibility of fraudulent reviews. To achieve this, we designed and implemented a restaurant review portal and a reservation system using real life data. Besides, we also designed a behavior tracking system to obtain key metrics such as time spent on each page, clicks, restaurant pages visited, and activities such as sorting by different parameters and finally, the details of what restaurant was chosen by the subject.

Randomized Experimental Design

To design our experimental scenarios, we consider three possible actions with which review portals can deal with fraudulent reviews from different levels: Filter, Inform, and Display. Review portals can make binary decisions at each of the three levels. i.e., on one extreme, a naive review portal does not have a fraud detection mechanism (Filter = 0), does not warn users about fraud (Inform = 0) and as a consequence of not having a fraud detection mechanism cannot display fraudulent reviews (Display = 0).

However, most of the review portals today have a fraud detection mechanism and differ mainly in whether they choose to Inform and Display the presence of the fraud to the users. Yelp, for example, has Filter = 1, Inform = 1 and Display = 1 which puts it on one end of the spectrum. On the other hand, Google does have a fraud detection mechanism but does not discuss about the possibility of fraud and thus we can assign Fraud = 1, Inform = 0 and Display = 0. There are some scenarios that are not possible due to the dependency among the three actions. For example, a review portal cannot display fraudulent reviews if it does not have a fraud detection mechanism in the first place. We exclude such scenarios in our experiment.

Scenario	**Referred as**	**Fraud Management Strategy**	**Currently Adopted by**	**Hypothetical**
A	Fraud Information Displayed	Fraudulent reviews were displayed but with a search cost of scrolling to the bottom of the page where they are linked to a different page	Yelp	No
B	Silent Approach	Do not display fraudulent reviews or acknowledge the presence of fraud or explain to the users the actions at the backend	Google	No
C	Fraud Information Displayed Prominently	Display the fraud reviews prominently in line with the normal reviews	-	Yes
D	Do not filter and Inform	Do not filter at the backend and inform users that no filtering was done at the backend	-	Yes
E	Filter and Inform	Filter fraud at the backend and inform users that filtering was done	-	Yes

Table 1. Description of Experimental Scenarios

In summary, given the real world feasibility and the focus of our research, we finalized our experimental design with a 5×1 mixed design. For between-subject design, we consider a total of five different scenarios. Table 1 illustrates detailed descriptions of each of the scenarios. We implemented each of the scenarios as a real-life decision making environment and assigned users randomly into one of the five scenarios. For within-subject design, we let each subject conduct two tasks for two different cities to control for any potential user-level unobservable and prior knowledge.

While each of these scenarios is interesting, first we focus on mainly Scenarios A & B, which are the major fraud management strategies adopted by the review portals. In Scenario A, fraudulent reviews are marked clearly and are provided to the user at the bottom of the review page and the users are free to look at them if they want to. This scenario corresponds to the fraud management strategy adopted by Yelp. Scenario B is the popular "Silent Approach" where the fraudulent reviews are removed and are not displayed to the users. This scenario is currently used by most of the review portals including Google.

Amazon Mechanical Turk (AMT)-Based User Behavioral Experiments

The website interface began with instructions about the experiment followed by a main page. The main page contained a list of fifty restaurants and a short summary of each restaurant's ratings, price and total number of reviews. This page also had options to sort by price and sort by rating. Once the subject clicked on a restaurant, she was taken to the restaurant's landing page where an elaborate description of the restaurant and reviews were present. The restaurant pages changed slightly in design, depending on the scenario the user was assigned to. Each user was exposed to only one scenario at a time.

The behavioral experiments in this paper were conducted on Amazon Mechanical Turk (AMT). AMT has been widely used in behavioral experiments and has been accepted in the literature as a standard platform to conduct randomized online user experiments. Paolacci et al. 2010 compared the results of standard decision making experiments conducted over AMT subjects with subjects from recruited from other online platforms and other subjects recruited offline from a university. The results differed only slightly quantitatively and did not differ qualitatively. Ipeirotis 2010 and Ghose et al. 2014 have conducted detailed user studies and shown that the AMT population is generally representative of the overall U.S. Internet population. Birnbaum 2000; Parkes et al. 2012; Suri and Watts 2011 have explored the logical consistency in decision making and other techniques to improve quality on AMT.

Incentive Based Quality Control

Paying attention to fraudulent reviews is a second level of investigation that happens in real life, only when there is a considerable investment at stake. We tackled this issue by informing the subjects that they would get a bonus ten times their participation wage if the restaurant they chose was among the top three restaurants adjudged by an internal panel unknown to the subjects. We also informed the subjects that the best way to get this bonus would be to pay attention and choose a restaurant like they would do in the real life. By making this highly incentive compatible, we ensured that the subjects internalized the stake and were incentivized enough to optimize their decision like they would in a real life scenario. We also told the subjects that the winning restaurants were not necessarily dependent on high ratings and all the information that they would need to make a good decision was present on the website.

Results

Impact of Displaying Fraudulent Reviews on User Search Behavior

We analyzed the experimental results at both the user level and the restaurant level. The corresponding results are provided in Table 2 and Table 3 (i.e., differences across scenarios that are statistically significant are highlighted in bold).

User Level Metric	**Scenario A (Fraud Information Displayed - Yelp)**	**Scenario B (Silent Approach - Google)**	**p-value**
Average Number of Restaurant Pages Visited per User	**5.641**	4.554	0.068
Average Clicks per User	**7.214**	5.256	0.008
Average Time Spent on a Restaurant Page per User (Minutes)	**4.531**	3.170	0.029
Average Fraud/Non-Fraud Ratio of the chosen restaurants	**0.125**	0.144	0.084

Table 2. Results from User Level Analysis

First, we compare the two scenarios that are currently used by the major review portals (i.e., Scenario A – Fraud Information Displayed *vs.* Scenario B – Silent Approach). Interestingly, our results from the user level show that users in Scenario A (Fraud Information Displayed) visited a significantly higher number of restaurants than in Scenario B (Silent Approach), which only filters fraudulent reviews at the backend but does not display or acknowledge them. This finding indicates that consumers, in their quest to optimize their choices, expand their choice set considerably when the fraudulent reviews are shown. This result is intriguing and it seems to suggest that users become more engaged during their decision making processes in Scenario A (Fraud Information Displayed). They tend to make their choices more carefully and to use their information at hand more rigorously when they are exposed to the suspected fraudulent information. This behavioral change can be beneficial to not only the users themselves, but also the products being visited, and moreover, the review portals in the long run. When users are exposed to a choice set with better variety, they are more likely to locate a product that fits their preferences better hence achieving higher satisfaction and consumer surplus (e.g., Brynjolfsson et al. 2003, Ghose et al. 2012). This fact can in turn lead to higher revenues for product search engines (Ghose et al. 2014).

Similarly, at the restaurant level, we find consistent trends. Our results from restaurant level show that restaurants visited in Scenario A (Fraud Information Displayed), on average attracted significantly higher number of visitors than the restaurants visited in Scenario B (Silent Approach) which does not display fraudulent review information. This finding provides further support that by displaying fraudulent information, review portals may facilitate user engagement by attracting more users to participate in product search.

Restaurant Level Metric	**Scenario A(Fraud Information Displayed - Yelp)**	**Scenario B (Silent Approach - Google)**	**p-value**
Average Number of Visitors to Restaurant Page	**3.870**	3.007	0.077
Average Number of Activities on Restaurant Page	**5.739**	3.908	0.032
Average Clicks on Restaurant Page	**7.339**	4.511	0.019
Average Time Spent on a Restaurant page (Minutes)	**1.091**	0.645	0.037

Table 3. Results from Restaurant Level Analysis

Meanwhile, we also noticed a similar trend from both user level and restaurant level analyses that users in Scenario A (Fraud Information Displayed) on average conducted significantly more clicks and spent more time per restaurant landing page than in Scenario B (Silent Approach), which does not display fraud reviews. This finding seems to suggest that users do behave differently when fraudulent reviews are displayed along with the non-fraudulent reviews. However, more time or activities spent by users may not necessarily mean users make better decisions. For example, users may spend more time browsing through the website simply because there is additional (fraudulent) information provided in Scenario A (Fraud Information Displayed), and they need to incur additional costs in processing such information. Therefore, to further examine the impact of displaying fraudulent reviews on user behavior, especially on the quality of user decision, we then look into the results of users' final choices (i.e., restaurants that are ultimately chosen by the users). We discuss the finding next.

Impact of Displaying Fraudulent Reviews on User Choice

To further examine the impact of displaying fraudulent reviews on user behavior, especially on the quality of user decision, we look into the restaurants that are ultimately chosen by the users. Interestingly, we found that users in Scenario A (Fraud Information Displayed) chose restaurants with a significantly lower average fraud/non-fraud ratio than users in Scenario B (Silent Approach). We provided the corresponding result in Row 5 in Table 2.

In our experiment, we knew each restaurant's history of fraud (like most review portals that have a fraud detection mechanism do). In Scenario A (Fraud Information Displayed), the users were aware of this information while in Scenario B (Silent Approach), the users were not aware of the information. In the former case, when fraud information was available, users evaluated a restaurant based on the history of fraud, and ended up choosing restaurants with a lower history of fraud. By not displaying the available historical fraud information to their users, the review portals are presenting honest businesses and the ones that solicit fraudulent reviews as equals before their consumers. Leveling the playing field between restaurants that resort to writing fake reviews versus the ones that do not, drastically lowers the reputation cost of being caught and does not carry any incentives for staying honest.

Discussion

In addition to comparing the two dominant scenarios that are adopted by most major review portals today, we also measured the same metrics in the sites with the other three hypothetical scenarios (i.e., Scenario C-Fraud Displayed Prominently in Line with Non-Fraud; Scenario D-Do not Filter and Inform; Scenario E-Filter and Inform). We find that Scenario A (Fraud Information Displayed) again outperforms any other hypothetical scenarios at user level in motivating user engagement by facilitating user search and expanding user choice set with a significantly higher variety. We also find similar trend at restaurant level that Scenario A (Fraud Information Displayed) also outperforms any other hypothetical scenarios in attracting significantly more unique visitors on site.

Overall, our results from Experiment I seem to indicate that filtering the fraud reviews silently at the backend is not enough. Instead, informing (warning) users about the potential existence of fraud is more effective in facilitating user decision making (regardless of whether the actual filtering is done or not). Most importantly, displaying this fraudulent information to users and reducing the information asymmetry is most effective. However, there is a key difference in performance depending on how review portals choose to display fraud information to their users. We find that displaying the fraud information prominently in line with the non-fraudulent reviews may not be best strategy either. In particular, Scenario A (Fraud Information Displayed) leads to more activities and more page views per restaurant compared to Scenario C (Fraud Displayed Prominently in Line with Non-Fraud). This result is intriguing. It suggests although displaying both fraud and non-fraud reviews can help improve user decision making, users tend to process these two types of information differently. In particular, users can incur different cognitive costs when processing these two types of information. Moreover, mixing them in line with each other can add significant information switching costs for the users and lead to a significant increase in the cognitive burden to the users. Therefore, our experimental findings indicate that not only displaying the fraudulent information is important in improving user decision making on the review portals, but also the way they are displayed is critical. In this paper, we propose a novel step into designing a single summarized score to help users reduce cognitive costs and process these two different types of information more efficiently.

All the metrics measured in Experiment I provide insights into the behavior of consumers when facing review portals with different fraud policies. Beyond the immediate user engagement and choice, a key factor that wins the loyalty of the consumers and makes them use a web portal regularly is the level of trust that they place on the web portal. Does displaying fraudulent reviews serve as a quality signal for the review platform? Do consumers perceive the displayed fraudulent reviews as a measure of efficiency in keeping fraudulent reviews at bay? Will any measure of indication that there was some kind of automatic filtering done at the backend help to increase the trust that the consumers place on the website? We address these questions in second randomized experiment.

Experiment II – Evaluating the Impact of Trust on Consumer Behavior

The primary objective of this experiment is to understand if displaying fraudulent increases the trust that the users place on the review portal. Does displaying fraudulent reviews indicate the review portal's ability to keep the review environment free from fraudulent reviews? We also attempt to study how the trust that the consumer places on the review portal change when the website attempts to display fraudulent reviews in a way that reduces the cognitive burden required to process the raw fraudulent information.

The importance of trust in e-commerce has been well documented in the marketing and computer science literature (e.g., Castelfranchi and Pedone n.d.; Meziane and Kasiran 2008; Patton and Jøsang 2004; Swearingen and Sinha 2001). Most the papers in the marketing literature use surveys to assess consumers' trust (e.g., yes/no or Likert scale responses). While this is an accepted technique, it is difficult to translate the amount of trust elicited through a survey into the actual amount of trust that a consumer would place in a web portal. Surveys are further limited in their ability to predict a consumer's trust in a system when there is no cost of making a bad decision or when a monetary stake involved.

This served as a motivation to design our second randomized experiment. We implemented a "trust game" to elicit the trust that user places on the review portal. A "trust game" is variety of a social dilemma game where the first mover aims to gain something by placing a certain amount of trust on the second mover. The first mover decides on how much to risk based on how much she would gain after the second mover makes her move.

The trust game is a technique is widely used in the Behavioral Economics literature. It evolves from the prisoner's dilemma games (Nash 1950) and the coordination aspect of prisoner's dilemma games has been extensively discussed in prior literature (e.g., Diekmann and Lindenberg 2001). Researchers in HCI have widely used coordination games to measure trust in Computer Mediated Communication (Bos et al. 2002; Davis et al. 2002; Jensen et al. 2000; Rocco 1998; Zheng et al. 2002). In our setting, the trust game allows us to make our subjects financially invest in their decisions, providing a more accurate measure of what they would do in their real life decision making process when there is a monetary risk involved in an E-commerce transaction.

This set up provides a good representation of the scenarios we are interested in testing. Here the consumers of the portal, as first movers, decide whether or not to trust a portal based on the information present on the portal. The consumers are placing themselves at risk - in real life of bad services, and in our experiment of considerable loss of money if they make a poor choice.

Randomized Experimental Design

The main goal of this randomized experiment is to examine the impact of displaying fraudulent reviews on user trust towards the review portal. Moreover, our observations from Experiment I show that not only displaying the fraudulent reviews is important, but also how the review portals choose to display this information is important. Therefore, in Experiment II, we apply a 4 × 2 experimental design. For within-subject design, again we allow for each subject to conduct tasks for two different cities to control for the potential subject-level unobservable. For between-subject design, we wanted to explore how we can effectively present the fraud information by focusing on the four scenarios as described in the Table 4.

In our experiment, we built a website where the subjects were presented with a home page (which we will refer to as the "betting page") that linked to the above four scenarios. The order in which the four scenarios were presented was randomized. The only difference between the scenarios was whether the scenario linked to fraudulent reviews at the bottom of the page, displayed a score on the restaurant page, or both, or neither. All the scenarios displayed non-fraudulent reviews.

Implementation of Trust Game

We ran our experiment on Amazon Mechanical Turk (AMT) from June 20 - 22, 2014. 109 subjects from the United States participated in the experiment. We paid the subjects a small participating fee and then gave the subjects $4 worth of virtual chips that they can use to bet on the sites they trusted the most. Note that the average wage for a ten-minute task is $0.80 (Ipeirotis 2010). Therefore $4 is a five-fold increase in the potential reward. In the instructions, we informed the workers that their bet should reflect the amount of trust they had on the site. Not betting was not an option. In other words, they could divide their $4 chips and invest them among the four sites based on how much they trusted each site. Since trust takes different meanings, in the instructions, we defined trust based on our context of a restaurant review portal as follows:

"Think of trustworthiness in the real world context - as a site you would believe gives you the best information and the one that you would want to use again and again if given the choice."

Note that to avoid potential user confirmation bias, we referred to the fraudulent activity summary score displayed in Scenarios C and D as a *"Review Quality Score"*, rather than using the word "trust" directly at a restaurant level. Sites displaying the Review Quality Score had a short description of what it meant, along with the score. The score that was displayed to the subjects was just the proportion of fraud to non-fraud reviews scaled to a score between 1 and 5. The main purpose of this score was to see if there was a major change in the amount of trust that the users placed on the review portal by displaying a decision heuristic. The short description of the review score was adapted from what Yelp uses under its non-recommended review page.

Similar to Experiment I, we designed a system that tracked users' activities throughout the website. In this way, we were able to confirm that all 109 subjects opened each of the websites and individual restaurant pages that contained the differences, before making their bets. In summary, the experiment had the two major scenarios from the previous experiment, one displaying fraudulent reviews along with the non-fraudulent reviews and one displaying just the non-fraudulent reviews. We added the "trust score" to the above two scenarios giving us totally four scenarios to work with.[1]

Scenario	Referred to as	Changes in Website
A	Fraud Information Displayed	Fraudulent reviews were linked to a separate page and this link present at the bottom of the restaurant page containing the non-fraudulent reviews
B	Silent Approach	Display only non-fraudulent reviews. Neither display the fraudulent reviews nor give any indication that any form of filtering was done at the backend
C	Fraud Information and Trust Score Displayed	Fraudulent reviews were linked to a separate page and this link present at the bottom of the restaurant page containing the non-fraudulent reviews. Summary of fraud in form of a score was provided
D	Only Trust Score Displayed (Along with Non-Fraudulent Reviews)	Summary of fraud in form of a score was provided along with the non-fraudulent reviews

Table 4. Description of Scenarios

Scenario	Description	Average amount of bets
A	Fraud Information Displayed	0.650
B	Silent Approach	0.562
C	Fraud Information and Trust Score Displayed	1.513
D	Only Trust Score Displayed (Along with Non-Fraudulent Reviews)	1.275
ANOVA significant at .001 level		

Table 5. Comparing Average Bets

Results

The results of the randomized experiment are presented in Table 5. A one-way ANOVA table shows that there was a statistically significant difference between each of these groups. This implies that subjects did evaluate each of these sites differently. They placed bets that were significantly different between the groups. This shows that subjects were cognizant of the changes in each of these sites. They related to some of the characteristics that they deemed trustworthy and that caused them to trust some sites more than others. We also note that if the subjects were indifferent between the options, they would, on average, have bet an equal $1 on all sites, which was not the case in our experiment.

Rather, on average, the subjects placed higher bets on sites that displayed the "Review Quality Score" versus other sites. A t-test between Scenario C and the rest of the sites shows that subjects placed more bets on the Scenario C, which displayed fraud reviews along with the Review Quality Score. In particular, the average bets were statistically significant for sites that contained the Review Quality Score versus those that did not. Between the two sites that had the Review Quality Score, subjects placed higher bets on the ones that displayed the fraudulent reviews along with the score. A t-test between Scenario A (Fraud review information was displayed along with non-fraudulent reviews) and Scenario D (just displaying the Review Quality Score along with non-fraudulent reviews) reveals that the subjects placed significantly higher bets on the later. Note that Scenario D is equivalent to the silent approach scenario but has the trust score displayed to the users.

[1] Note that we renamed the trust score to "Review Quality Score" in the experiment to avoid potential user confirmation bias.

This shows that user display a clear preference for a website that reduces their cognitive burden in processing fraud information in form of a score as compared to parsing the raw fraud information from the fraudulent reviews. What is also interesting is that the currently popular Silent Approach in Scenario B was considered to be the least trust-worthy. If review portals do not display fraudulent reviews fearing the loss in trust that the consumers place on the website, our experiment suggests the contrary. By displaying the fraudulent reviews and by decreasing the cognitive cost in processing the fraudulent reviews, the website actually increases consumers' trust on the review portal.

Discussion

Review portals today aim to help consumers with product evaluation, which can potentially lead to follow-up purchases. The business model revolves around the consumer trust on the user-generated content. Products such as restaurants are experience goods and consumers' trust in the portal depends on the accuracy of the information provided. From Experiment II, it is evident that consumers' trust in the review portal increases when they notice a score that provides a summary of fraudulent activities and the quality of the customer reviews.

Moreover, our subjects trusted sites that displayed the fraudulent reviews along with credibility more than sites that just displayed the review trust score (i.e., Review Quality Score). The link to these fraudulent reviews was accessible at the bottom of the page. Interestingly, we notice that while most of our subjects did not actually open the fraudulent review page, the mere fact that it was accessible to them seemed to increase their trust in the site. This shows that such review trust score also serves as a measure of a review portal's efficiency in cleaning out fraudulent reviews. The trust score seems to reassure consumers that there is some mechanism at the backend that looks out for fraudulent reviews and keeps the platform's reviews accurate. Our subjects also displayed a clear preference for a platform that provides a summarized heuristic and reduces the cognitive burden in understanding fraud compared to a platform that just displays all the fraud reviews.

The other implication of empowering consumers with detailed information about suspected reviews, both in the form of reviews and a trust score, is that it creates a system of reputation and increases the cost to commit fraud. DellaVigna and La Ferrara 2010 and Mayzlin et al. 2012 argue how a reputational mechanism serves as a huge deterrent in committing future fraud. The current practice of weeding out fraudulent reviews at the backend does not deter dishonest businesses from writing fraudulent reviews, as the reputational costs of being caught are non-existent. This fosters an environment to commit more fraud in hopes of getting past the fraud detection mechanism.

Therefore, the proposed trust score can not only be used to reduce the cognitive burden in processing fraud information but also can be considered as a first step in establishing a system of accountability to reduce information asymmetry to the users. This score can also be designed in a way to increase the costs of dishonesty and prevent businesses from committing fraud. In the next section we discuss how we can design such a trust score.

Designing the Trust Score

In this section of our paper we suggest a novel method to develop this score. In order to develop the trust score we first develop a probabilistic model that robustly detects fraudulent reviews from non-fraudulent reviews. Then, we apply this model to develop a trust score that measures the expected unbiased reviewer rating by imposing a penalty on potential review fraud. The basic intention is to create a score that takes the original rating, which is generally an average over all non-fraudulent reviews, and adjust this rating depending on the quantity of fraudulent reviews and the probability that we can attach to every review being fraudulent. Therefore, a review portal that already has a fraud detection mechanism in place can implement the trust score derived from our model and a review portal that doesn't have a fraud detection mechanism can implement our complete model to detect fraud reviews with good accuracy and also display a trust score.

Model

Model Setting: Define a review to be a "Positive" review if the rating is greater than 3, and a "Negative" review if the rating ≤ 3. For a review i of restaurant j in week t, we observe the following probabilities from our data:

- $\Pr(Positive_{ij})$, $\Pr(Negative_{ij})$: Probability that restaurant *j* receives a positive or a negative review *i*;
- $\Pr(Fraud_{ij})$: Probability that review i is fraud for restaurant *j*;
- $\Pr(Positive_{ij} \mid Fraud_{ij})$, $\Pr(Negative_{ij} \mid Fraud_{ij})$: Conditional on restaurant j receiving a fraud review i, the probability that the review is positive or negative;
- $\Pr(Positive_{ij} \mid Not\_Fraud_{ij})$, $\Pr(Negative_{ij} \mid Not\_Fraud_{ij})$: Conditional on restaurant j receiving a non fraud review i, the probability that the review is positive or negative;

Meanwhile, according to the definitions, we know that:

$$\Pr(Negative_{ij}) = 1 - \Pr(Positive_{ij})$$

$$\Pr(Negative_{ij} \mid Fraud_{ij}) = 1 - \Pr(Positive_{ij} \mid Fraud_{ij})$$

$$\Pr(Negative_{ij} \mid Fraud_{ij}) = 1 - \Pr(Positive_{ij} \mid Fraud_{ij})$$

Overall Likelihood: For restaurant j, the joint probability of observing a total of N_j reviews with F_j out of the N_j reviews being fraud and NF_j out of the N_j reviews being not fraud, including F_j^+ positive frauds, $F_j - F_j^+$ negative frauds, NF_j^+ positive non-fraud reviews and $NF_j - NF_j^+$ negative non-fraud reviews can be written as:

$$\begin{aligned}
&\Pr(N_j, F_j, F_j^+) \\
&= \prod_{i=1}^{F_{jt}} \Pr(Fraud_{ijt}) \cdot \prod_{i=1}^{N_{jt}-F_{jt}} \left[1 - \Pr(Fraud_{ijt})\right] \prod_{i=1}^{F_j^+} \Pr(Positive_{ij} \mid Fraud_{ij}) \cdot \\
&\prod_{i=1}^{F_j - F_j^+} \left[1 - \Pr(Positive_{ij} \mid Fraud_{ij})\right] \cdot \prod_{i=1}^{F_j^+} \Pr(Positive_{ij} \mid Not\_Fraud_{ij}) \cdot \prod_{i=1}^{F_j - F_j^+} \left[1 - \Pr(Positive_{ij} \mid Not\_Fraud_{ij})\right]
\end{aligned}$$

To understand how different review, restaurant, and user level characteristics may affect this joint probability, we further model the individual probabilities as functions of these different factors.

Modeling Fraud Probability: The probability restaurant *j* will receive a fraud review *i* as a function of restaurant, review and user characteristics

$$\Pr(Fraud_{ij}) = \alpha_0 + \alpha_1 \text{Price}_j + X_j \beta_1 + T_{ij} \beta_2 + R_{ij} \beta_3 + \varepsilon_{ij}$$

where X is a set of restaurant level characteristics, T is a set of review level characteristics and R is the set of reviewer characteristics·

Modeling the Rating Distribution of Fraud/NonFraud: The conditional probability of a restaurant j to receive a positive fraud review and positive truthful review i is

$$\Pr(Positive_{ij} \mid Fraud_{ij}) = \gamma_0 + \gamma_1 \text{Price}_j + X_j \phi_1 + T_{ij} \phi_2 + R_{ij} \phi_3 + \upsilon_{ij}$$

$$\Pr(Positive_{ij} \mid Not\_Fraud_{ij}) = \omega_0 + \omega_1 \text{Price}_j + X_J \mu_1 + T_{ij} \mu_2 + R_{ij} + \tau_{ij}$$

Note that by definition $Pr(\text{Postive}|\text{Fraud}_{ij})$ and $Pr(\text{Postive}|\text{Not\_Fraud}_{ij})$ are conditional probabilities and they are independent from (Fraud_{ij}) which allows us to estimate our model. We used Maximum Likelihood Estimator (MLE) for estimation.

Designing the Trust Score

Our goal is to design a trust score that can reveal the quality and truthfulness of the online reviews. Note that one unique feature of our proposed trust score is that we treat positive fraud reviews and negative fraud reviews differently. More specifically, reviewers who post positive fraud reviews (i.e., promotional reviews) and those who post negative fraud reviews (i.e., malicious reviews) are likely to have different motivations. We propose to design a novel trust score that can penalize these two types of fraud activities differently.

- **Self-Promotion** - When products receive positive fraud (i.e., promotional) reviews we should take into account the "probability of dishonesty" as one additional dimension of measuring product quality. We aim to increase the risk of "self-promotion."
- **Bad-Mouth** – When products receive negative fraud (i.e., malicious reviews) the focal business (i.e., who suffers from the malicious reviews potentially from its competitors) should be compensated based on the probability of being bad-mouthed. In other words, we aim to increase the risk of maligning other business unfairly.

To achieve the above goals, we propose a new unbiased rating mechanism by incorporating these perspectives of risks from the two different types of fraud reviews.

$$OverallRating_j\_FraudAware = OriginalRating_j - \frac{1}{N_j^+}\sum_{i}^{i\in N_j^+} \Pr(Fraud_{ij} \mid Positive_{ij}) * Rating_i$$

$$+\frac{1}{N_j^-}\sum_{i}^{i\in N_j^-} \Pr(Fraud_{ij} \mid Negative_{ij}) * Rating_i$$

Therefore, the proposed trust score changes the original rating of the restaurant depending on the proportion of fraudulent reviews, which is observed, and on the probability with which each review can be classified as fraudulent which needs to be calculated. To calculate this, we need to first derive the predicted posterior probability of fraud conditional on observing an upcoming review as being positive or negative. We can then infer the posterior probabilities as follows:

$$\Pr(Fraud_{ij} \mid Positive_{ij}) = \frac{\Pr(Positive_{ij} \mid Fraud_{ij})\Pr(Fraud_{ij})}{\Pr(Positive_{ij})} \propto \Pr(Positive_{ij} \mid Fraud_{ij})\Pr(Fraud_{ij})$$

$$\Pr(Fraud_{ij} \mid Negative_{ij}) = \frac{\Pr(Negative_{ij} \mid Fraud_{ij})\Pr(Fraud_{ij})}{\Pr(Negative_{ij})} \propto \Pr(Negative_{ij} \mid Fraud_{ij})\Pr(Fraud_{ij})$$

When a restaurant receives a new positive or negative review, we are able to predict the probability of the review being fraud based on the observed characteristics. Once we have the predicted conditional fraud probabilities, we can then leverage them to designing the final score.

Data

Our dataset to perform MLE, consists of 283,830 fraud and non-fraud reviews obtained from 982 restaurants in San Francisco on Yelp.com. Yelp is a useful platform for our purposes because it has a fraud detection mechanism in place that marks fraud reviews separately from non-fraud reviews. This dataset takes into account all the reviews written after June 2012 up until December 2014. Each observation in our dataset is a review. We consider all reviews with a rating greater than 3 stars to be positive, and all reviews with a rating less than or equal to 3 stars to be negative. We also indicate whether a review is fraud based on the fraud/not-fraud status given by Yelp.com.

The overall rating of the restaurant is provided as the average over non-fraudulent ratings. We also use the historical fraud data of the restaurant in form of the total number of non-fraud reviews written and total number of fraud reviews detected in our model. We created a facilities score out of 14 for each restaurant looking at the kind amenities that they provide such as Wi-Fi, presence of TV, etc.

The dataset also contains reviewer level information like number of reviews written, number of friends and whether the user had an elite status. From the raw text of the reviews, we extracted review level information such as the number of words, number of sentences, nouns, verbs, adjectives, nouns and verbs. We also performed sentiment analysis on reviews and extracted the subjectivity on each of the reviews. A highly subjective review will have a score of 1 and a highly objective review will have a score of zero. A subjective review tends to draw from personal experiences while an objective review tends to be unbiased and factual.

Model Evaluation

Predictive Performance of the Model

We use the Receiver Operating Characteristic (ROC) Curve to compare the performance of different models using different subsets of data to account for the potential variation in detecting the empirical boundary of fraud. ROC curve is widely used in Computer Science and Machine Learning communities (Bradley 1997; Pencina et al. 2008) to understand the performance of binary prediction models when the discrimination decision thresholds are varied. The value points on the ROC curves are measured in terms of the "True Positive Rate" and the "False Positive Rate. The top left corner (i.e. the (0,1) point in the graph) denotes perfect classification. The diagonal indicated by the dotted line denotes a random guess model. Intuitively, the closer a point is located to the top left corner, the higher is the corresponding model performance. In our case, True Negative refers to non-fraud reviews identified correctly as non-fraud and False Positive refers to non-fraud reviews incorrectly identified as fraud.

Figure 3 shows the ROC curves of different models using different subsets of data. The solid black line in Figure 3 illustrates the ROC curve of our main MLE model. We found that our proposed MLE model with the entire restaurant, user and review features showed an accuracy of 75% with an 84% true positive rate and a 27% false positive rate at the 0.5 decision threshold. Interestingly, the ROC curve indicates that our model performs best when the decision threshold shifts to 0.47, leading to an 86% true positive rate and a 30% false positive rate.

In addition, to explore how different levels of features may affect the predictive performance, we considered similar models with different sets of features in Figure 3. In particular, we tried three different models: 1) MLE model but with only restaurant related features; 2) MLE model but with only review related features; 3) MLE model but with only user related features. For 1 and 2, we found that the overall accuracy decreases dramatically (as illustrated in Figure 3 by the long dash and dotted lines). Model 3 (as indicated by the dot-dash line in Figure 3) performs almost as well as our original MLE model with all the features, which indicates that most of the predictive power of the model comes from the user level features of the model.

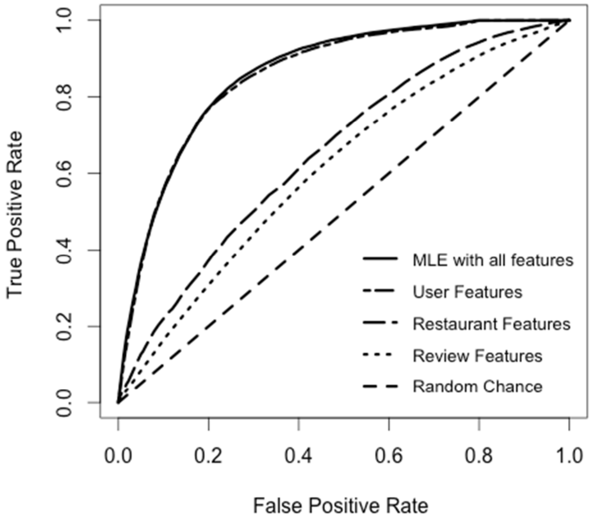


Figure 3. Model Comparison Using Different Levels of Data

To evaluate the robustness of our model prediction, we also compared our results with several popular models used in Machine Learning using both in-sample and out-of-sample predictions. We found a significant improvement in the accuracy of prediction between our model and that of Naıve Bayes. Support Vector Machine (SVM) is one of the most widely used Machine Learning models and performs very well even on data that is not linearly separable. We found a negligible difference in the accuracy of prediction between our model and that of an SVM. There was also a negligible difference in accuracy between our model and that of Decision Trees. These findings indicate that our model is robust to over fitting, and can be potentially generalized to other settings.

Discussion

We emphasize that our observational data analysis aims to propose a novel approach of leveraging the predicted fraud probability as *an intermediate stage* towards designing a final "trust score" that reflects the quality and truthfulness of online reviews. All our results were obtained based on the publicly observable characteristics of restaurants, reviews and users on Yelp. Review portals can usually access even richer information at the backend and therefore the prediction accuracy using our model should be considered only as a lower bound. Recent work (e.g., Liu 2010 and Mukherjee and Liu 2012) suggests that the prediction of fraud can be improved greatly by observing group spam characteristics. These include reviews written from similar IPs, the propensity to write multiple reviews by the same user on various restaurants in a suspiciously short time frame or an unexpectedly high number of similar reviews for a restaurant in a short time frame are potentially important features in distinguishing fraud reviews from non-fraudulent ones. These can be easily added as extra features in our model.

Conclusions and Implications

In this paper, we combine randomized user experiments and observational data analysis using statistical learning method to understand the best strategy in dealing with fraudulent reviews after they are detected. Our final results allow us to design a novel trust score that can be displayed on the review portal summarizing the history of fraudulent activities. This trust score can reveal the quality and truthfulness of online reviews.

Our main findings are the following: 1) Our experimental results show that by displaying fraudulent review information on review portals, the users tend to click more, engage more, and demonstrate a higher trust towards the review portal. Interestingly, we find that instead of mixing the fraudulent and non-fraudulent reviews in line with each other for display, it is more efficient to display them on separate pages along with an overall adjusted rating that corrects for bias from the potentially fraudulent reviews. It is critical for review portals to not only display the fraudulent information but also display it in an effective way to reduce consumer cognitive cost. 2) We observe that by displaying fraudulent reviews, users tend to choose those restaurants that have a lower historical fraud probability. In our observational data analysis, we find that this historical information of higher fraud to non-fraud ratio is a strong indicator of a restaurant's tendency to commit fraud in the future. Therefore, by not displaying this information, review portals endanger their users in making a sub-optimal decision. Also, by not displaying fraudulent reviews, the review portals essentially level out the playing field and thereby lower the cost of committing fraud. 3) Finally, we find that users trust a review portal more when it displays a summary heuristic of fraudulent activities in the form of a trust score. Therefore, we propose to design a novel trust score that can decrease the cognitive burden and also penalize the two fraud activities differently. In summary, our results show that the optimal strategy for a review portal in dealing with fraudulent reviews is not only to detect them but also to display these reviews to users. These reviews should be displayed in a way that it reduces the cognitive burden involved in processing the information.

Though our study focused on restaurants, we expect that our results will hold in high-risk settings like health care and is an area for future research. Displaying fraudulent reviews will also raise awareness in this age of increasing cybercrime and will caution people not to make high cost decisions based only upon reviews. The onus on the review portal is not just in creating robust fraud detection mechanisms, but more importantly, in displaying this information to its users in the form of a simplified heuristic. The model that we proposed in this paper is one such mechanism that will not only detect fraudulent reviews but also help in creating a trust score.

Finally, we note that our analysis has several limitations. First, our experiments were conducted on Amazon Mechanical Turk (AMT). While the behavioral research literature (Birnbaum 2000; Paolacci et al. 2010; Parkes et al. 2012; Suri and Watts 2011) has consistently shown that experiments conducted on AMT provide equivalent results to the experiments conducted in a lab, ideally we should have been able to observe real users in a real-world setting. While randomization of the experiments helps reduce the concern of heterogeneity and selection, as in any experimental work, it is difficult to evaluate how users would behave under real world conditions and what they would perceive as the inherent risk in making a bad decision while looking at reviews. Second, we acknowledge that that trust is not a one-dimensional quantity and in our experiment we only considered the setting where the reviewer has no prior knowledge about the portal. Excluding these limitations, we believe that our paper helps future research by helping technology-based companies better understand dynamics of the cat-and-mouse game of online fraudulent information.

References

Ahluwalia, R., and Shiv, B. 1997. “Special Session Summary the Effects of Negative Information in the Political and Marketing Arenas: Exceptions to the Negativity Effect,” *Advances in consumer research* (14), pp. 213–217.

Akoglu, L., Chandy, R., and Faloutsos, C. 2013. “Opinion Fraud Detection in Online Reviews by Network Effects.,” *ICWSM*.

Anderson, M., and Magruder, J. 2012. “Learning from the crowd: Regression discontinuity estimates of the effects of an online review database*,” *The Economic Journal* (122:563), Wiley Online Library, pp. 957–989.

Archak, N., Ghose, A., and Ipeirotis, P. G. 2011. “Deriving the pricing power of product features by mining consumer reviews,” *Management Science* (57:8), INFORMS, pp. 1485–1509.

Ba, S., and Pavlou, P. A. 2002. “Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior,” *MIS quarterly*, The Society for Information Management and The Management Information Systems Research Center of the University of Minnesota, and The Association for Information Systems, pp. 243–268.

Birnbaum, M. H. 2000. “Psychological experiments on the Internet,” Elsevier.

Bos, N., Olson, J., Gergle, D., Olson, G., and Wright, Z. 2002. “Effects of four computer-mediated communications channels on trust development,” in *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 135–140.

Bradley, A. P. 1997. “The use of the area under the ROC curve in the evaluation of machine learning algorithms,” *Pattern recognition* (30:7), Pergamon, pp. 1145–1159.

Brynjolfsson, E., Y. Hu, M. D. Smith. 2003. "Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers," *Management Science,* 49(11), pp. 1580-1596.

Castelfranchi, C., and Pedone, R. (n.d.). “A review on trust in information technology, 1999,” *Online http://alfebiite. ee. ic. ac. uk/templates/papers. htm.*

Chatterjee, P. 2001. “Online reviews: do consumers use them?,” ACR.

Chen, Y., and Xie, J. 2008. “Online consumer review: Word-of-mouth as a new element of marketing communication mix,” *Management Science* (54:3), INFORMS, pp. 477–491.

Chevalier, J. A., and Mayzlin, D. 2006. “The effect of word of mouth on sales: Online book reviews,” *Journal of marketing research* (43:3), American Marketing Association, pp. 345–354.

Davis, J. P., Farnham, S., and Jensen, C. 2002. “Decreasing online’bad'behavior,” in *CHI’02 Extended Abstracts on Human Factors in Computing Systems*, pp. 718–719.

Dellarocas, C. 2003. “The digitization of word of mouth: Promise and challenges of online feedback mechanisms,” *Management science* (49:10), INFORMS, pp. 1407–1424.

DellaVigna, S., and La Ferrara, E. 2010. “Detecting illegal arms trade,” *American Economic Journal: Economic Policy* (2:4), American Economic Association, pp. 26–57.

Diekmann, A., and Lindenberg, S. 2001. “Sociological aspects of cooperation,” *International Encyclopedia of the Social & Behavioral Sciences. New York: Elsevier Science*, Citeseer.

Gartner. 2013. “Gartner,” (available at http://www.gartner.com/newsroom/id/2161315).

Ghose, A., Ipeirotis, P. G., and Li, B. 2012. “Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content,” *Marketing Science* (31:3), INFORMS, pp. 493–520.

Ghose, A., Ipeirotis, P. G., and Li, B. 2014. "Examining the Impact of Ranking on Consumer Behavior and Search Engine Revenue," *Management Science* 60(7), INFORMS, pp. 1632–1654.

Ipeirotis, P. G. 2010. “Analyzing the amazon mechanical turk marketplace,” *XRDS: Crossroads, The ACM Magazine for Students* (17:2), ACM, pp. 16–21.

Jensen, C., Farnham, S. D., Drucker, S. M., and Kollock, P. 2000. “The effect of communication modality on cooperation in online environments,” in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 470–477.

Jindal, N., and Liu, B. 2008. “Opinion spam and analysis,” in *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pp. 219–230.

Jindal, N., Liu, B., and Lim, E.-P. 2010. “Finding unusual review patterns using unexpected rules,” in *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 1549–1552.

Kanouse, D. E. 1984. “Explaining negativity biases in evaluation and choice behavior: Theory and research,” *Advances in consumer research* (11:1), Association for Consumer Research Provo, UT, pp. 703–708.

Lascu, D.-N., and Zinkhan, G. 1999. “Consumer conformity: review and applications for marketing theory and practice,” *Journal of Marketing Theory and Practice*, Association of Marketing Theory and Practice, pp. 1–12.

Lim, E.-P., Nguyen, V.-A., Jindal, N., Liu, B., and Lauw, H. W. 2010. “Detecting product review spammers using rating behaviors,” in *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 939–948.

Liu, B. 2010. “Sentiment analysis and subjectivity,” *Handbook of natural language processing* (2), Chapman & Hall, pp. 627–666.

Luca, M., and Zervas, G. 2013. “Fake it till you make it: Reputation, competition, and Yelp review fraud,” *Harvard Business School NOM Unit Working Paper* (14-006).

Mayzlin, D., Dover, Y., and Chevalier, J. A. 2012. “Promotional reviews: An empirical investigation of online review manipulation,” National Bureau of Economic Research.

Meziane, F., and Kasiran, M. K. 2008. “Evaluating trust in electronic commerce: a study based on the information provided on merchants’ websites,” *Journal of the Operational Research Society* (59:4), Nature Publishing Group, pp. 464–472.

Milgrom, P., and Roberts, J. 1986. “Price and advertising signals of product quality,” *The Journal of Political Economy*, The University of Chicago Press, pp. 796–821.

Mukherjee, A., and Liu, B. 2012. “Aspect extraction through semi-supervised modeling,” in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pp. 339–348.

Nash, J. F. 1950. “Equilibrium points in n-person games,” *Proceedings of the national academy of sciences* (36:1), pp. 48–49.

Nelson, P. 1974. “Advertising as information,” *The journal of political economy*, JSTOR, pp. 729–754.

Netzer, O., Feldman, R., Goldenberg, J., and Fresko, M. 2012. “Mine your own business: Market-structure surveillance through text mining,” *Marketing Science* (31:3), INFORMS, pp. 521–543.

Ott, M., Choi, Y., Cardie, C., and Hancock, J. T. 2011. “Finding deceptive opinion spam by any stretch of the imagination,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 309–319.

Paolacci, G., Chandler, J., and Ipeirotis, P. G. 2010. “Running experiments on amazon mechanical turk,” *Judgment and Decision making* (5:5), pp. 411–419.

Parkes, D. C., Mao, A., Chen, Y., Gajos, K. Z., Procaccia, A., and Zhang, H. 2012. “Turkserver: Enabling synchronous and longitudinal online experiments,” in *Proceedings of the Fourth Workshop on Human Computation (HCOMP’12)*.

Patton, M. A., and Jøsang, A. 2004. “Technologies for trust in electronic commerce,” *Electronic Commerce Research* (4:1-2), Kluwer Academic Publishers, pp. 9–21.

Pencina, M. J., D’Agostino, R. B., and Vasan, R. S. 2008. “Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond,” *Statistics in medicine* (27:2), John Wiley & Sons, Ltd., pp. 157–172.

Resnick, P., and Zeckhauser, R. 2002. “Trust among strangers in internet transactions: Empirical analysis of ebay’s reputation system,” *The Economics of the Internet and E-commerce* (11:2), pp. 23–25.

Riegelsberger, J., Sasse, M. A., and McCarthy, J. D. 2003. “Shiny happy people building trust?: photos on e-commerce websites and consumer trust,” in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 121–128.

Rocco, E. 1998. “Trust breaks down in electronic contexts but can be repaired by some initial face-to-face contact,” in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 496–502.

Suri, S., and Watts, D. J. 2011. “Cooperation and contagion in web-based, networked public goods experiments,” *PLoS One* (6:3), Public Library of Science, p. e16836.

Swearingen, K., and Sinha, R. 2001. “Beyond algorithms: An HCI perspective on recommender systems,” in *ACM SIGIR 2001 Workshop on Recommender Systems* (Vol. 13), pp. 1–11.

Wu, G., Greene, D., Smyth, B., and Cunningham, P. 2010. “Distortion as a validation criterion in the identification of suspicious reviews,” in *Proceedings of the First Workshop on Social Media Analytics*, pp. 10–13.

Yatani, K., Novati, M., Trusty, A., and Truong, K. N. 2011. "Review spotlight: a user interface for summarizing user-generated reviews using adjective-noun word pairs," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1541–1550.

Yoo, K.-H., and Gretzel, U. 2011. "Influence of personality on travel-related consumer-generated media creation," *Computers in Human Behavior* (27:2), Pergamon, pp. 609–621.

Zheng, J., Veinott, E., Bos, N., Olson, J. S., and Olson, G. M. 2002. "Trust without touch: jumpstarting long-distance trust with initial social activities," in *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 141–146.