Abstract

Online platforms are prone to abuse and manipulations from strategic parties. For example, social media and review websites suffer from sentiment manipulations, manifested in the form of opinion spam and fake reviews. The consequence of such manipulations is the deterioration of information quality as well as loss in consumer welfare. Applying the economic concept of rational expectation equilibrium (REE), we explore the impact of manipulation on consumer welfare in a Twitter-like environment. We argue that the REE outcome can be decomposed into a firm-centric effect and a rational expectation effect, and the relative strengths of these effects determine the final level of manipulation. Our preliminary empirical study on movie tweets sheds light on the reliability of sentiment analysis, and contributes to our understanding of strategic manipulation. We argue that appropriate verification strategies increase the cost to manipulate, and can consequently dissuade firms from engaging in strategic manipulations.

1. Introduction

Much economic activity involves the understanding of consumers’ preferences and the subsequent recommendations of products of interest, both of which are instrumental to product-selling firms’ performances. Thanks to their popularity and ability to reach diverse demographic groups, internet platforms have established themselves as powerhouses where consumers seeking information can interact among themselves as well as with sellers, and sellers can actively identify target consumers and channel their advertisements accordingly. These online platforms include e-commerce website such as Amazon; social media sites such as Facebook, Twitter, and Foursquare; and recommendation and review websites such as Yelp, TripAdvisor, and Expedia. What these platforms have in common is the ability for consumers to voice their opinions, and for sellers to inform potential consumers of the quality of their products or services in various ways. These platforms also establish a better communication channel between sellers and buyers. Both consumers and sellers can benefit from active participation on these platforms because they provide consumers with much more detailed information on products, while providing firms the opportunity to reach out to target consumers.

However, popular platforms are prone to abuse and manipulation from strategic parties, and these manipulations can cause dire consequences. For example, to increase product visibility, producers might want to manipulate platform data by adding positive sentiments themselves, so the sentiment analysis results would be more favorable to their products; platforms, trying to increase visitor traffic and encourage more user participation, do not necessarily want to eradicate such producer-generated data pollution.

Generally speaking, manipulation is a result of lack of awareness, absence of verification mechanisms, platforms’ incentive, or the nature of the chosen business model. We explore the relationship between the level of sentiment manipulation and consumers’ actions under rational expectation equilibrium in the current study. Our findings suggest that, when firms are low quality producers, the level of manipulation will be higher than the case when firms are high quality producers. We also examine how the level of manipulation would be different if consumers were aware of the presence of fraudulent information, and we find that the negative effect of manipulation is the largest when consumers are naive and not aware of any manipulation.

Our results reflect the combination of two effects that determine the equilibrium outcome: a firm-centric effect and a rational expectation effect. The firm-centric effect drives the firm to pursue a high level of manipulation, since more fake information would make consumers more likely to believe that a given product is of high quality; in contrast, the rational expectation effect dissuades the firm from manipulating too much, since the more manipulation, the more consumers will discount any information they receive. The relative strength of the two effects determines the total effect. We also examine the effect of competition on firm
manipulation, and, similarly, the equilibrium outcome can be decomposed into a competition effect and a rational expectation effect. The competition effect induces firms to manipulate more in the presence of rivals, while the rational expectation effect discourages manipulation. Our preliminary empirical study examines the reliability of Twitter sentiment in the context of movie tweets, and we find that Twitter sentiment shows an unusual pattern compared to critic ratings, which prompts us to suspect the existence of manipulation. In addition, there are websites that specialize in selling Twitter followers and Twitter accounts, and virtually anyone is able to use the acquired accounts and followers to manipulate sentiment.

Although researchers in computer science have made efforts to detect and filter out malicious behaviors, advertisers and platforms’ economic incentives are more difficult to change. In addition to modeling consumers’ response to the presence of spam, we also discuss how verification strategies and better control of message content can potentially curb the manipulation issue. Our research contributes to the literature on sentiment aggregation by formulating a rational expectation equilibrium framework to model the firm’s incentive to manipulate, and by analyzing the resulting effects on sentiment analysis and consumer welfare. Our analyses also suggest that practitioners should be cautious when conducting sentiment analysis on user generated content, as the designs of these platforms make them susceptible to strategic manipulation.

2. Literature review

User generated content (UGC) and social media data have been used in all aspects of decision making processes [2, 10, 12]. Despite the effectiveness of UGC and social media data in improving business decisions, several studies have empirically shown the existence of widespread manipulation practices on these sites. Mayzlin et al. [9] examine the prevalence of difficult-to-detect fake reviews on popular review websites. More specifically, they use a difference in differences approach to look at how hotel characteristics and ownership structure affect the level of review manipulation, which consists of posting positive reviews for one’s own business and manufacturing negative reviews for competitors, on travel websites Expedia.com and TripAdvisor.com. Luca and Zervas [6] investigate the presence of restaurant review fraud on another review website, Yelp.com. They find that positive review fraud is related to reputational concerns, while negative review fraud is more likely due to competitions. Anderson and Simester [1] offer a different perspective on the nature of deceptive reviews. Using a dataset from a private apparel retailer, they find that, in addition to firms’ strategic behaviors, customers without clear financial incentives to manipulate product ratings might still write reviews on products they did not purchase.

Dellarocas [3] constructs a theoretical model on firms’ manipulative behaviors. He shows that manipulations could be beneficial to consumers if firms’ manipulation strategies are monotonically increasing in their true qualities. He also shows that, under certain threshold conditions, firms would actually benefit if manipulation were not possible. Mayzlin [8] examines marketers’ incentives to generate anonymous promotional messages online. Using a game theoretic model, her results show that, contrary to traditional advertising strategies, firms producing low quality products would engage in more promotional chat than those producing high quality products. This is because high quality product benefits from positive word of mouth which substitutes for advertising, while low quality product does not. Our paper is distinct from these two prior studies in the following ways: (1) we formally define rational expectation equilibria in the context of sentiment analysis, and explicitly compare the differences induced by having rational-expectation-forming consumers and naive consumers; (2) Mayzlin [8] assumes that some portions of the consumers are “informed” in the sense that they know exactly the quality of the product. In our setting, we do not require that the consumers know the product quality; instead, the informed consumers in our setup only have to know that a proportion of messages are posted strategically by the firm itself. More specifically, we adopt the rational expectation equilibrium framework, popularized by the influential works of Sargent and Wallace [13] and Lucas [7], to examine firms’ incentive to generate opinion spam in a Twitter-like social broadcasting environment. The use of a rational expectation equilibrium in modeling both firms and consumers’ equilibrium behaviors can also be found in Su and Zhang [14].

3. Model setup

3.1 Single-firm case

---

1 [http://twitteraddicts.com/buy-twitter-followers/] sells Twitter followers; LIVESON ([http://liveson.org/]) is a web service specializing in social afterlife: This website will keep producing tweets for users even if they have passed away.
We formulate the firm’s profit maximization problem where the firm chooses the optimal level of manipulation in response to platform consumers’ purchasing decisions. We consider the case where a single firm manipulates, as well as the case where there are multiple firms competing with each other, all of them able to manipulate the sentiment. We will then discuss how the rational expectation equilibrium framework can provide us with very different predictions from the case where we assume consumers to be naive.

To more closely integrate our theoretical discussion with the later empirical analysis, we choose to base our analysis on a Twitter-like environment, and therefore we use the term “Twitter” to refer to the opinion platform, and “tweets” to refer to the contents on these platforms. We emphasize that our theoretical results can be applied to any opinion platform or review website that features user generated contents.

Formally speaking, let \( m \) be the number of fake positive tweets that the firm decides to post on Twitter to maximize its expected profit, where the cost to manipulate is given by a cost function \( a(m)^b \). \( a \) and \( b \) are some known constants. For simplicity, we assume that the firm posts all of its tweets before any consumer posts her tweets. These consumers arrive sequentially, and the quality of any given product can be either good, \( G \), or bad, \( B \), i.e. \( q \in \{G,B\} \).\(^2\) The consumers receive a private signal, \( s_i \), independently and identically drawn from a Bernoulli distribution, \( p \), and this signal can be either high, \( H \), or low, \( L \), with the high (low) signal meaning the product is more (less) likely to be of good quality. We assume a common prior on the probability that the product is of good quality: \( \Pr(G) = \Pr(B) = 1/2 \), and we also assume informative signals: \( \Pr(H|G) = \Pr(L|B) = 3/4 \); \( \Pr(L|G) = \Pr(H|B) = 1/4 \). Then for the \( i \)-th consumer, the combination of tweets she would see on Twitter includes both the previous consumers’ tweets, \( s_1, \ldots, s_{i-1} \), her own tweet, \( s_i \), together with the firm’s \( m \) fake positive tweets. We assume the consumer is aware of the possibility that certain tweets might have been manipulated, but she is not able to distinguish genuine positive tweets from fake positive tweets, presumably because the firm uses different aliases to post. Therefore, a rational consumer would discount the number of positive tweets by some discount factor, \( f \), which corresponds to her belief of the level of genuine tweets among all positive tweets. Since we only consider rational expectation equilibria, this discount factor \( f \) must also equal the proportion of genuine positive tweets among all positive tweets posted on Twitter. Based on its belief on the probability of any consumer receiving an \( H \) signal, the firm’s problem is to choose the number of fake positive tweets to post. Let the firm’s belief of any consumer receiving an \( H \) signal be some probability \( p \), and let the total number of consumers be \( n \). Then the expected number of genuine positive tweets is \( n \cdot p \). Depending on the value of \( m \), the expected proportion of genuine positive tweets and, equivalently, the consumer’s discount factor, \( f \), can be expressed as \( f = np/(np+m) \). For simplicity, we assume the consumer only discounts the positive tweets she sees, while she trusts her private information and thus never discounts her own signal. Also define \( h_i \) as \( h_i = \sum_{j=1}^{i-1} I(s_j = 1) \). Notice that, aside from her private signal, the \( i \)-th consumer sees \((m+h_{i-1})\) positive tweets and cannot distinguish the firm’s fake positive tweets from genuine positive tweets. At a rational expectation equilibrium, she rationally and correctly expects the proportion of genuine tweets, and adopts the discount factor \( f \). So the effective level of positive tweets is now \( f \cdot (m+h_{i-1}) = \frac{n \cdot p}{n \cdot p + m} \cdot (m+h_{i-1}) \), and the effective level of negative tweets is \((i-1) \cdot h_{i-1} \cdot (-1)\), where, without loss of generality, we have first converted each \( H \) signal to have the numerical value 1, then discounted it from 1 to \( f \); \( 0 \leq f \leq 1 \); and also converted each \( L \) signal to -1; \( h_{i-1} \) follows a binomial distribution. Denote the quantity \( f \cdot (m+h_{i-1}) + ((i-1) \cdot h_{i-1} \cdot (-1)) \) as \( S_i^* \). Then the consumer’s decision can be specified as follows:

\[
D_i = \begin{cases} 1 & \text{if } S_i^* + s_i > 0 \\ 1 & \text{if } [S_i^* + s_i = 0] \land [s_i = 1] \\ 0 & \text{otherwise} \end{cases}
\]

We provide a formal definition of a rational expectation equilibrium involving one strategic firm and \( n \) rational consumers.

1. A rational expectation equilibrium with one strategic firm and \( n \) rational consumers consists of \((f, \{D_i\}_{i=1,\ldots, n}, m^*)\) which satisfies

\[(i)\] \( D_i = I(S_i^* + s_i > 0) + I(S_i^* + s_i = 0) \land [s_i = 1] \),

\[(ii)\] \( m^* = \arg \max_m \pi(p, m(p)) = \arg \max_m \sum_{j=1}^{n} E[D_j[p] - a(m)^b] \),

\[(iii)\] \( f = np/(np + m^*) \), where \( S_i^* \) is defined as \( f \cdot (m+h_{i-1}) + ((i-1) \cdot h_{i-1}) \cdot (-1) \); \( \pi(\cdot) \) is the firm’s

\(^2\) We recognize that quality is highly subjective. We choose to use a simple binary variable to represent quality in order to focus our theoretical discussion on the characterization of a rational expectation equilibrium. More elaborate measures of quality, such as introducing both vertical and horizontal differentiation, can be a future research direction.
expected profit function:
\[ E[D_i | p] = \Pr\{S_i^* + S_i > 0\} + \Pr\{S_i^* + S_i = 0\} \land \{s_i = 1\} \].

Notice that condition (i) is from the rational consumer’s optimal decision \( D_i \) based on her belief of the proportion of genuine tweets, \( f \); condition (ii) is from the strategic firm’s optimal choice of its manipulation level; condition (iii) ensures consistency. It is worth noting that in a rational expectation equilibrium, the consumers are always able to correctly expect the number of genuine tweets and fake positive tweets, and thus the discount factor, regardless of the sizes of these numbers.

3.2 Single-firm results

In this section we describe our simulation results and compare how the rational expectation equilibrium model provides different estimates from the model of naive consumers, where we define naive consumers to be those who do not discount any sentiments. For the first set of simulations, we set the number of consumers to be 30; the cost function, \( a(m)^b \), is specified with \( a = 0.1 \) and \( b = 2 \), i.e., convex (quadratic) cost function. We note that the simulation results are robust to different values of \( a \) and number of consumers. This cost function specification implies that the marginal cost of producing fake tweets increases in the manipulation level, potentially because it would become increasingly difficult to avoid detection once the number of fake tweets is large, and thus more costly for the firm to manipulate.

We vary the probability that the product is of high quality from 0.3 to 0.7. The manipulation results based on different probability levels allow us to understand how the product quality affects the firm’s incentive to manipulate. Figure 1 shows the different levels of manipulation as we vary the probability of the product being of good quality from 0.3 to 0.7. We can see that, in general, the higher the probability of the product being of good quality, the less manipulation the firm would generate. This result is consistent with the findings from Mayzlin (2006) where she shows that firms that produce low-quality products spend more on promotional chats. It is worth noting that there is a significant difference between the naive and the rational cases where the firm would try to manipulate more when the users are naïve.

To calculate consumer welfare, we assume the utility of buying a good product is 1, that of a bad product is -1, and the utility of not buying any product is 0. We also assume that consumers are risk neutral. Figure 2 plots the levels of consumer welfare for both the naive and rational cases, against varying levels of product quality. This figure shows that the difference
in welfare level between naive and rational consumers is the largest when the product quality is low. In other words, naive consumers suffer the most when facing low-quality products, while there is no significant difference between naive and rational consumers when the product quality is sufficiently high. Moreover, firms in general would benefit more when consumers are naive, and the difference is most pronounced when the product quality is low, as shown in Figure 3. This means that, as long as consumers are naive, firms do not necessarily need to increase the quality of the products; instead, they could possibly rely on posting fake positive tweets in order to attract consumers to buy their products.

Figure 4. Manipulation (a=0.1; b=1)

We also simulate the case where the cost function is linear instead of convex. The results are similar to the convex cost case, except for the manipulation level when the consumers are rational. Comparing Figure 1 and 4 we can see that, when the marginal cost of manipulation is increasing, the firm manipulates more in the naive case; with a constant marginal cost of manipulation, the firm manipulates more in the rational case. This might be because that, since the cost of manipulating is relatively small, the firm would try to manipulate more when facing rational consumers in order to compensate for the discounting of positive tweets that those consumers would engage in, so the resulting positive sentiment could still be favorable for the firm. Note that the type of marginal cost faced by manipulative firms can be determined by the verification and detection mechanisms employed by the platforms. For example, platforms without strict verification policies such as Twitter and TripAdvisor might fall in the category of constant marginal cost, which means that, in the rational consumer setting, firms would engage in more manipulative behaviors on these platforms. Other platforms, in contrast, have mechanisms that can better ensure the integrity of user generated contents: Yelp has a review filter that hides suspicious reviews (Luca and Zervas 2013); only customers who have reserved and visited restaurants are allowed to submit ratings on the restaurant reservation platform OpenTable. This type of platforms uses verification mechanisms to deter abusive use and manipulation, which can be modeled as having an increasing marginal cost, and results shown in Figure 1, Figure 2, and Figure 3 would apply. It is worth noting that, without taking into account the consumer’s forming rational expectation with regard to firms’ manipulation, we would have underestimated the level of manipulation, had the cost function been linear.

Figure 5. Welfare (a=0.1; b=1)

To provide some intuitions for these results, we consider two forces that affect the equilibrium outcomes. Firstly, there is a firm-centric effect which encourages the firm to manipulate more when the consumers are rational. This is because rational consumers expect some level of manipulation, so the firm would have to exert extra efforts in order to counter the consumer’s discounting. On the other hand, when consumers are naive, the firm can more easily convince them to purchase its product, and hence they would not have to manipulate as much. However, there is another effect, which we coined the rational expectation effect, that induces the consumers to engage in more sentiment discounting the more likely the firm wants to manipulate, and thus the less effective the manipulation would be. Therefore, when the cost of manipulation is high, the rational expectation effect will predict a decrease in manipulation. Overall speaking, a rational expectation
equilibrium reflects both of the effects, and the results depend on the relative strengths of them. In the convex cost case, since manipulation is relatively costly, the rational expectation effect outweighs the firm-centric effect, so we observe a lower level of manipulation in the presence of rational consumers compared with the naive setting, as illustrated in Figure 1. Similarly, since the cost of manipulation is relatively cheap in the linear cost case, the firm-centric effect dominates, and we observe more manipulation in the rational setting than that in the naive setting, as shown in Figure 4.

![Figure 6. Profit (a=0.1; b=1)](image)

### 3.3 Multiple-firm case

We can extend our discussion of the rational expectation equilibrium to the case where there are multiple strategic firms and $n$ rational consumers. We discuss the two-firm case in this paper, since adding more firms would be a straightforward generalization of the two-firm results.

To simplify the model, we only consider the case where the firm would only produce fake positive messages regarding its own product. We defer the possibility of fake negative messages to future studies. First we suppose that all players in this model possess common priors on the probability of product $A$ receiving an $H$ signal, denoted as $p_A$; and, similarly, the probability of product $B$ receiving an $H$ signal, $p_B$.

Without loss of generality, we assume $p_A > p_B$. Firm $A$ and firm $B$ sell similar products and the consumer can only purchase at most one firm’s product. We further assume that both firms know there are a total of $n$ consumers in the market. Since consumers resort to Twitter sentiment for decision making, the firms have incentives to strategically post fake positive tweets. In a rational expectation equilibrium, the consumer does not know the actual level of manipulations firm $A$ and firm $B$ would pursue. Instead, she possesses some beliefs over the proportion of positively manipulated tweets among tweets related to both firms, denoted $f^A$ and $f^B$, respectively. We assume that all consumers share the same $f^A$ and $f^B$. Firm $A$ decides the number of fake tweets it will post, $m^A$, given its beliefs on firm $B$’s manipulation level, $R^B$. Similarly, firm $B$ would post $m^B$ fake positive tweets to promote its own product, based on its belief of firm $A$’s manipulation level, $R^A$. To simplify this complex model, we assume that firms only post fake positive tweets about their own products, and do not post fake negative tweets about the competitor’s product.

A rational expectation equilibrium in this context means that all consumers choose a purchase strategy, $D_j$, to maximize their expected utilities, while the firms maximize their expected profit, $\pi^A$ and $\pi^B$, given the consumer’s purchasing strategy and competitors’ manipulation level. For the $i$-th consumer, we assume that she observes a total of $k^A$ tweets on product $A$, i.e. $\sigma^A_1 \ldots \sigma^A_{k^A}$, and $k^B$ tweets on product $B$, i.e. $\sigma^B_1 \ldots \sigma^B_{k^B}$, both excluding her own signals, $s^A_i$ and $s^B_i$. Since she cannot distinguish genuine tweets from fake tweets, we use $\phi^A(i)$ to express product $A$’s discounted average sentiment that she can observe, where

$$
\phi^A(i) = \left[ f^A \left( \frac{1}{k^A} \sum_{j=1}^{k^A} \sigma^A_j = 1 \right) \right] + s^A_i;
$$

Product $B$’s discounted average sentiment observed by consumer $i$ is defined similarly. Firm $A$’s objective is to maximize its expected profit by choosing a manipulation level, $m^A$, given its belief on firm $B$’s manipulation strategy, $R^B$. Its maximization problem is denoted by

$$
m^A( R^B ) = \arg\max_{m^A} \pi^A( p_A, p_B, R^B, m^A );
$$

firm $B$’s maximization problem is defined similarly. Therefore, from firm $A$’s point of view, it can predict both $\phi^A(i)$ and $\phi^B(i)$ by the following equations:
\[ \phi_{A}^i(i) = \frac{n \cdot p_A}{n \cdot p_A + m^A} \left( h_{i-1}^A + m^A \right) + (i-1-h_{i-1}^A \cdot (-1) + s_{i}^A; \right. \\
\phi_{B}^i(i) = \frac{n \cdot p_B}{n \cdot p_B + m^B} \left( h_{i-1}^B + m^B \right) + (i-1-h_{i-1}^B \cdot (-1) + s_{i}^B, \right. \\
\]

where \( h_{i-1}^A = \sum_{j=1}^{i-1} (s_j^A = 1) \) and \( s_j^A \) are genuine tweets; the superscript \( A \) refers to product \( A \)'s sentiment; the subscript [4] refers to firm \( A 's knowledge about \( \phi^A(i) \). Notice that the first term on the right hand side of the \( \phi_{A}^i(i) \) equation consists of two parts: (1) the total number of positive tweets regarding product \( A \) that consumer \( i \) observes, which is itself a summation of all organic positive tweets up to the consumer immediately before her, \( h_{i-1}^A \), and product \( A \)'s own manipulative tweets, \( m^A \), and (2) the discount factor, \( n \cdot p_A \left( n \cdot p_A + m^A \right) \) which is the expected proportion of genuine positive tweets, \( n \cdot p_A \), among all positive tweets, \( n \cdot p_A + m^A \); the second term on the right hand side of \( \phi_{A}^i(i) \) equation is the number of negative tweets up to consumer \( i-1 \), which is the total number of organic tweets excluding positive organic tweets, \( i-1-h_{i-1}^A \), multiplied by the weight -1; the last term \( s_{i}^A \) is consumer \( i \)'s own signal, which is not discounted since we assume consumers' own signals are not manipulated. Note that firm \( B \)'s prediction of \( \phi^A(i) \) and \( \phi^B(i) \) can be defined similarly. Then, we have the following rational expectation equilibrium definition: 

2 A rational expectation equilibrium for two strategic firms and \( n \) rational consumers consists of \( (m^A, m^B, \{D_i\}_{i=1,...,n}, f^A, f^B, R^A, R^B) \) which satisfies 

(i) \( D_i(f^A, f^B) = \Psi(\phi^A(i), \phi^B(i)) \) 

(ii) \( m^A(R^B) = \argmax_{m^A} \pi^A(p_A, p_B, R^B, m^A) \)

\[ = \argmax_{m^A} \sum_{i=1}^{n} \mathbb{E}(D_i = A|p_A, p_B, R^B, m^A) - \alpha(m^A)^{b}, \]

(iii) \( m^B(R^A) = \argmax_{m^B} \pi^B(p_A, p_B, R^A, m^B) \)

\[ = \argmax_{m^B} \sum_{i=1}^{n} \mathbb{E}(D_i = B|p_A, p_B, R^A, m^B) - \alpha(m^B)^{b}, \]

(iv) \( m^A = R^A \frac{n \cdot p_A \cdot (1-f^A)}{f_A}, \)

\( m^B = R^B \frac{n \cdot p_B \cdot (1-f^B)}{f_B}. \)

where \( p_A \) and \( p_B \) are common priors for the two firm's products, respectively;

\[ \Psi(\phi^A(i), \phi^B(i)) = \begin{cases} 
0 & \text{if } \phi^A(i) < 0 \land \phi^B(i) < 0 \\
1 & \text{if } \phi^A(i) > 0 \land \phi^B(i) > 0 \\
\frac{1}{2} & \text{otherwise}
\end{cases}. \]

Notice that condition (i) specifies consumers' decision rules; condition (ii) and (iii) ensure that the level of manipulation that firm \( A \) and firm \( B \) attempt is equal to their respective profit-maximizing levels; condition (iv) ensures the rational expectation equilibrium in the sense that the resulting manipulation level of each firm is equivalent to its competitor's expectation and the consumer's expectation of its manipulation level. The consumer's decision rule, \( \Psi(\phi^A(i), \phi^B(i)) \), makes use of the following principles: (1) the consumer would always choose the product with a higher discounted tweet sentiment, as long as that sentiment is positive. If both products' sentiment levels are negative, then the consumer will not purchase either one; (2) the consumer's own private signal is assumed to be always genuine, and serves as a tie breaker when the number of discounted positive tweets equals the number of negative tweets; (3) if a tie persists after applying the previous two principles, then we assume that the consumer will take a Bernoulli(1/2) random draw to decide which product to purchase.

3.4 Multiple-firm results

Figure 7 and 8 plot the simulation results from rational expectation equilibria with two competing firms. Each plot represents three sets of firm types, indexed by "High", "Medium", and "Low". The "High" case is specified with the two firms both having high probabilities of producing good quality products, with the two probabilities being Pr(quality=Good) =
0.7 and 0.6, respectively. Intuitively, we can understand these probabilities as different firm types, with the higher probability firm producing a better quality product, and the lower probability firm producing a lower quality product. Table 1 lists all combinations of firm types that are used in our simulation. The results in Figure 7 correspond to the changes in profit, sales, and manipulation levels of Firm A when it competes with Firm B, compared with the case where Firm A is a monopoly, with the associated firm product qualities listed in Table 1. Note that the focal firm in Figure 7 is Firm A, namely the firm producing the higher quality product among the two competing firms; Figure 8 illustrates the effect of competition on Firm B, the firm producing the inferior product among the two competing firms.

Table 1. Two-firm cases product quality levels

<table>
<thead>
<tr>
<th>Pr[Quality=Good]</th>
<th>Firm A</th>
<th>Firm B</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>0.7</td>
<td>0.6</td>
</tr>
<tr>
<td>Medium</td>
<td>0.6</td>
<td>0.4</td>
</tr>
<tr>
<td>Low</td>
<td>0.4</td>
<td>0.3</td>
</tr>
</tbody>
</table>

We also point out that the effect of competition is measured by comparing the level of manipulation, consumer welfare, firm sales, and firm profits, in the two-firm case, with those in the monopoly setting, holding the focal firm’s product quality constant. For example, the “High” case in Figure 7 represents the effect of competition on a firm with a product quality of 0.7, when facing a competitor that produces a product of a lower quality, 0.6. The “Medium” and “Low” cases defined in Table 1 can be interpreted similarly. As mentioned earlier, we simulate both the case where this high quality firm faces no competition, and the case where it faces competition from a lower quality firm, and compare these two cases to quantify the effects of competition. Similarly, the “High” case in Figure 8 illustrates the effect of competition on a firm which produces a product with a quality level of 0.6, when competing with a firm that produces a better product with a quality level of 0.7. Detailed discussion on the effect of competition on firm manipulation is available upon request.

4. Empirical evidence

In our preliminary empirical discussion we examine how the sentiment measure reflected on Twitter differs from a benchmark quality measure, in the context of the movie industry. More specifically, we randomly select a movie, *The Perks of Being a Wallflower* (2012), and collect all tweets related to this movie using the hashtag, #theperksofbeingawallflower, starting 60 days before the release date, until 60 days after the release date. We train a Naive Bayes Classifier with unigram features based on the corpus described in Go et al. [5], and use this classifier to measure the probability of any given tweet being positive. We then use this probability as a measure of polarity. Figure 9 shows the daily average sentiment scores of this movie across this 120-day window.

We also collect critic ratings from the movie review site, Rotten Tomatoes, to construct a benchmark measure of movie quality. Critic ratings are reliable quality sources because the non-anonymous feature of
reviews and thus a high reputational cost would deter the critics from engaging in manipulative behaviors. For promotional purposes, movie studios usually invite movie critics to view the movie prior to the official release, and therefore some reviews will be published before the release date. A reasonable assumption is that, critics reviews should only reflect post-consumption evaluation of the film, since reviewers should only review a movie after watching it. Therefore, no critic reviews should be affected by whether or not the movie has been released, whereas the consumer sentiment on Twitter might experience a release shock near the release date, because prior to movie release, it is much easier for movie studios to manipulate movie sentiment due to the low number of existing consumer tweets at this stage, so any added positive tweets can more easily render the overall sentiment positive. In fact, we will argue that, since different consumers form different expectations about a given movie, on average the difference between the sentiment of any given tweet and the actual quality of the movie should be very small. This means that what the release shock captures can be understood as movie studios' strategic sentiment manipulation.

We use a regression discontinuity design to examine the release shock described above, and the results of the graphical analysis are shown in Figures 9 and 10, where we plot the patterns of Twitter sentiment and critic ratings over time. We can see from Figure 9 that there is indeed a discontinuity on the release day in the Twitter sentiment of *The Perks of Being A Wallflower*; on the contrary, Figure 10 shows that there is no significant discontinuity in the critic ratings. These patterns in Twitter sentiment and critic ratings are consistent with our expectation. Notice that the local linear fit is generated via local linear regression. See Fan and Gijbels [4] for details.

While refraining from making any causal statements, we do point out the observation that Twitter sentiment exhibits interesting pattern around the release date, whereas critic ratings on Rotten Tomatoes remain consistent. This observation leads us to suspect that some kind of manipulation might be involved in the observed Twitter sentiment. We emphasize that further research is needed in order to demonstrate the existence of sentiment manipulation.

![Figure 9. Twitter sentiment of “The Perks of Being a Wallflower” with local linear fit](image)

5. Concluding remark and future direction

In this paper we studied the effect of manipulation on consumer welfare, and the effect of competition on the firm’s manipulation decision, both in the rational expectation equilibrium framework. We emphasized the importance of recognizing the existence of strategic manipulation, because researchers as well as practitioners have been collecting and analyzing tremendous amount of social and opinion platform data to conduct sentiment analysis, often without explicitly adjusting for manipulated sentiments. Our results suggested that the equilibrium outcomes of manipulation level can be decomposed into a firm-centric effect and a rational expectation effect. When marginal cost to manipulate is increasing, the rational expectation effect dominates the firm-centric effect, and the firm will consequently manipulate less. We also considered the effect of competition on the firm’s incentive to manipulate. We found that, when the firm’s product quality is low, it is likely that the rational expectation effect will dominate the
competition effect, which would discourage the firm from manipulating. Our empirical study also showed an unusual pattern in Twitter sentiment, which prompted us to suspect the existence of manipulation.

We recognize several limitations in the current research: (1) we only considered fake positive messages in our theoretical model. However, in reality, firms often post negative messages about competitors to lower their competitors’ sentiment. This type of fake negative message attacks will further complicate the firm’s optimal manipulation strategy, and should be carefully studied in future research; (2) we were unable to empirically estimate the relative proportions of rational and naive consumers on the platforms; (3) in the empirical analysis, factors such as the movie studio’s advertising expenditure, fan activity, and other movie characteristics were not considered in the analysis; (4) we only used one simple classifier to generate the sentiment score. As a future extension, we plan to incorporate the possibility of posting negative messages, and to conduct a more thorough empirical study on manipulative behaviors on opinion platform, with multiple more advanced classifiers. Moving forward, Twitter, as well as other review platforms, must address the spamming and verification issues in order to avoid the danger of losing values and relevance in the years to come. Other online platforms should also reassess their susceptibility to manipulative behaviors, and find ways to maintain their credibility for a sustainable development.

6. References


