Fake Reviews: The Malicious Perspective

Theodoros Lappas

Boston University, Computer Science Dept.
tlappas@cs.bu.edu

Abstract. Product reviews have been the focus of numerous research efforts. In particular, the problem of identifying fake reviews has recently attracted significant interest. Writing fake reviews is a form of attack, performed to purposefully harm or boost an item's reputation. The effective identification of such reviews is a fundamental problem that affects the performance of virtually every application based on review corpora. While recent work has explored different aspects of the problem, no effort has been done to view the problem from the attacker's perspective. In this work, we perform an analysis that emulates an actual attack on a real review corpus. We discuss different attack strategies, as well as the various contributing factors that determine the attack's *impact*. These factors determine, among others, the *authenticity* of fake review, evaluated based on its linguistic features and its ability to blend in with the rest of the corpus. Our analysis and experimental evaluation provide interesting findings on the nature of fake reviews and the vulnerability of online review-corpora.

1 Introduction

Item reviews are nowadays found in abundance in a plethora of websites on the World Wide Web. The reviewed items include a wide range of different products (e.g. electronics, books, movies) and service providers (e.g. restaurants, hotels). Users write reviews to rate a particular item and express their opinion on the item's various attributes. The accumulated volume of reviews on the item then serves as a valuable source of information for a potential customer. Not surprisingly, it has been shown that reviews have a tremendous effect on the popularity of an item [5, 8, 21, 23].

Before the establishment of the web and the modern e-commerce model, reviews were authored by eponymous experts, typically considered authorities in their respective fields. In today's online world, however, anyone can author and publish an anonymous review on a major review-hosting site such as *Amazon* or *Yelp*. The option of anonymity, combined with the major impact of reviews on the users' purchase decisions, make review corpora a prime target for malicious attacks. An attack on the review corpus of an item includes the injection of fake reviews in order to harm or boost the item's reputation. A fake review does not provide an accurate representation of the item's quality and can thus be misleading to users. In addition, false information can have a negative effect on any application that is defined in the context of the review corpus (e.g. review search [15], selection [24],summarization [27]).

The detection of fake reviews is a well-recognized problem that has attracted significant interest from the research community [10, 11, 16–18, 25]. While previous work has made advances in the identification of fake reviews, no effort has been made to approach the problem from the point of view of the attacker. In this paper, we present an analysis of attack scenarios on review corpora. As we demonstrate in our work, the design and implementation of such an attack is a non-trivial task. In fact, the attacker needs to address a number of relevant questions, such as:

- *What information should I include in a fake review?*
- *How can I maximize the impact of the attack on the reputation of the target item?*
- *How can I make a fake review to appear as authentic as possible?*

Contribution: Our work is the first to formalize the parameters that need to be considered by the aspiring attacker. By unraveling the attack-process, our findings can be used toward the immunization of review corpora and the creation of enhanced methods for the detection of fake reviews. Our corpus-sensitivity measure can be used by a review-hosting website in order to properly allocate its limited resources for fraud detection. Further, a merchant (e.g. a restaurant owner) can be more vigilant given the knowledge that the corpus of reviews on their item is highly sensitive to a potential attack.

2 Writing Fake Reviews

A successful attack needs to take into consideration the factors that are most influential to its success. In our work, we identify and examine the following two factors:

1. **Authenticity:** In order for a fake review to be convincing, and thus less likely to be detected, it needs to be as authentic-looking as possible.
2. **Impact:** The injected fake review needs to be written in a way that maximizes the (positive or negative) impact on the target's reputation.

In the following sections, we formalize and discuss each of these two factors in detail.

2.1 Authenticity

First, we identify the three factors that affect the authenticity of a fake review:

1. **Stealth** measures the ability of the review to blend in with the corpus.
2. **Coherence** evaluates whether the assigned rating is in accordance with the opinions expressed in the review's text.
3. **Readability** depends on the structural features of a review's text and captures how easy it is to parse.

Next, we discuss and motivate these factors via the use of appropriate examples.

[Stealth]: Consider a restaurant evaluated on four attributes: *food quality, service quality, parking options* and *price*. For these four attributes, we assume the following opinions are observed in the review corpus:

- food quality: 200 positive, 5 negative
- service quality: 100 positive, 90 negative
- parking options: 5 positive, 1 negative
- price: 200 positive, 0 negative

Now consider the following fake review, assigning the minimum 1-star rating (out of 5) to this restaurant.

Review 1: *I'm never going to this restaurant again. First of all, finding a place to park around this place is impossible. It took us forever. Things only got worse after we actually got there. Our server was rude and entirely unprofessional. He pretty much ignored our requests and made us feel unwanted. On top of that, the food was terrible and seriously overpriced. One of the worst eating experiences of my life.*

This review will obviously have a high impact, if injected to the corpus. Observe that, in addition to assigning the minimum possible rating, the review also expresses a negative opinion on *all* the attributes. However, when one considers the distribution of opinions in the existing corpus, the review stands out as suspicious. In particular, the reviewer expresses (among others) highly negative opinions on the food and the price. This contradicts the very strong positive consensus that is observed in the corpus on these particular attributes, and hints that the reviewer is malicious or at least biased. Even if a review-management system does not eliminate this review, it can consider decreasing its visibility, in order to avoid presenting such questionable information to its users.

The above example clearly illustrates the following: in order to avoid having the fake review blacklisted or demoted, the attacker needs to consider the distribution of opinions in the existing corpus. The stronger the consensus on a given attribute, the more suspicious it appears if the fake review contradicts it. Further, the more positive and less negative opinions there are on an attribute, the stronger the positive consensus on an that attribute. Similarly for a negative consensus. To capture this intuition, we formalize the *stealth* of a fake review r in the context of a review-corpus R as follows:

$$stealth(r,R) = \frac{\sum\limits_{\alpha \in \mathcal{A}_r} |\{r' : r' \in R, r[\alpha] = r'[\alpha]\}|}{\sum\limits_{\alpha \in \mathcal{A}_r} P(\alpha,R) + N(\alpha,R)} \tag{1}$$

where $P(\alpha,R)$ is the number of positive opinions on attribute α in R, and $N(\alpha,R)$ is the respective number of negative opinions. Also, $\mathcal{A}_r$ is the set of attributes evaluated in r, and $r[\alpha] \in \{-1,+1\}$ returns the polarity of the opinion expressed in review r on attribute α. A value of +1 (-1) is returned for a positive (negative) polarity. The numerator represents the number of reviews that agree with r on attribute α, for every attribute $\alpha \in \mathcal{A}_r$. The denominator is equal to the sum of all the opinions expressed on all the attributes in $\mathcal{A}_r$. Obviously, the lower the fraction, the more suspicious the review appears.

[Coherence]: While the above formalization captures the concept of stealth, it disregards the review's *rating*. In the context of the previous example, a savvy attacker can easily write a stealthy review, such as the following:

Review 1: *Great food and very reasonably priced. That aside, I was not happy with our server. She was impolite and always in a hurry. Parking was also a bit of a pain, it took us a while to find a spot.*

After making sure that the opinions expressed in the review are in accordance with the rest of the corpus (see the distribution in the previous example), the attacker can then assign a very low rating (e.g. 1 star) in order to have a major impact on any rating-based ranking function. However, such a review would not be *coherent*, since the opinions expressed in the text are mixed and cannot justify such a low rating. This example motivates the concept of *coherence* for reviews. Intuitively, opinions expressed in a review should be in accordance with not only the other reviews in the corpus, but also with its own rating. For simplicity, we organize ratings into 3 mutually-exclusive groups: positive (+1), negative (-1) and neutral (0). For example, for the 5-star scale, the respective groups could be $\{4,5\}$, $\{1,2\}$ and $\{3\}$. We then formalize *coherence* as follows:

$$coh(r) = \begin{cases} P(r)/|r| & rating(r) = +1 \\ N(r)/|r| & rating(r) = -1 \\ \frac{(|r|-|P(r)-N(r)|)}{|r|} & rating(r) = 0 \end{cases} \tag{2}$$

where $P(r)$ and $N(r)$ return the number positive and negative opinions expressed in r, respectively. If the assigned rating is positive (negative), then coherence is the probability that a randomly chosen opinion expressed in the review is positive (negative). For neutral ratings (zero), the ideal scenario occurs when the review contains an equal number of positive and negative opinions. In other words, in the ideal scenario, every negative opinion is balanced by a positive one. Thus, for a neutral rating, coherence is the fraction of matched opinions.

[Readability]: The final factor that contributes to a review's authenticity is its *readability*. This is a well-studied concept that has motivated a significant body of relevant work [7]. In our methodology, we apply the well-known Flesh-Reading Ease (FRE) formula [6] for the evaluation of readability. Our choice is motivated by the popularity of this measure (incorporated in systems like Microsoft Office Word, WordPerfect and WordPro), as well as its successful application in the domain of reviews in relevant work [19, 14]. FRE is formally defined as follows:

$$FRE(r) = 206.835 - 1.015 \times \frac{words(r)}{sents(r)} - 84.6 \times \frac{syllables(r)}{words(r)} \tag{3}$$

where *words(r), sents(r) and syllables(r)* return the number of words, sentences and syllables in r, respectively. The Flesch Reading Ease yields numbers from 0 to 100, expressing the range from *very difficult* to *very easy*.

2.2 Impact

An attack on a review corpus aims at manipulating the reputation of the target-item. Toward an appropriate formalization of *impact*, we need a review-based measure that objectively captures the *reputation* concept. On a high level, we assume the existence of a reputation-evaluation function $f(\cdot)$ which, given the set of reviews on an item, assigns a score that captures the item's popularity, as encoded in the review corpus. Most modern review-hosting sites employ the average rating, observed over all the reviews in the corpus. We include this definition in our analysis, formalized as:

$$f_1(R) = \frac{\sum_{r \in R} rating(r)}{|R|}. \tag{4}$$

where R is the corpus of reviews for a given item and $rating(r)$ returns the rating for a given review r. The rating can be defined in the popular 5-star scale or on any other ordinal scale. Despite its popularity, the average rating often fails to deliver an accurate evaluation of the item's quality, since it treats every item as a single unary entity. In practice, however, reviewers evaluate an item based on the quality of its numerous *attributes*. For example, a digital camera can be evaluated based on its battery life, the picture-quality, the design, the price etc. The individual opinions that focus on particular attributes can be much more informative than the assigned star rating. Consider the following 3-star review on an Italian restaurant, taken from a major review-hosting site:

Example 1. This place has a lot of personality. The ambiance is great and the greeters were terrific despite the fact that we had a huge party coming in. They seated us at a table that was comfortable even though we were packed in pretty tightly. We started off with bread and some garlic oil concoction that one of my friends had ordered. It was delicious. I then ordered the Lasagna which was meatless or so I was told and I must say was quite mouthwatering. The sauce was tasty and the pasta itself was very good. I know the rest of my party was very pleased with their meal so I think I will definitely come back if not for just the bread and the oil dipping thingy. Overall, good experience, nothing over the top but nothing subpar either. My only complaint would be the service which i thought was slow and they didn't refill my drink once even though I had asked for it.

If one were to ignore the text of the review and focus only on the rating, the 3 stars assigned by the reviewer imply an impression of a mediocre restaurant. However, the text reveals attributes worthy of praises like “delicious”, “mouthwatering”, “tasty” and “very pleased”. In fact, the only complaint of the reviewer was related to the service. For a potential customer who is mostly interested in the quality of the food, this review would be considered extremely favorable.

Motivated by the above discussion, we consider an *attribute-based* evaluation function. Thankfully, a long line of related research has provided us with methods for extracting opinions from reviews [4, 20]. Specifically, given a review r, we can identify the attributes of the item that the reviewer comments on, as well as the polarity (positive/negative) of each expressed opinion. We can thus define the following reputation function for an item, given its corpus of reviews R and the set of its attributes $\mathcal{A}$:

$$f_2(R) = \frac{\sum\limits_{\alpha \in \mathcal{A}} (P(\alpha, R) - N(\alpha, R))}{|R|} \tag{5}$$

Here, $\mathcal{A}$ is the set of the item's attributes. $P(\alpha, R)(N(\alpha, R))$ is the number of positive (negative)opinions expressed in the entire corpus R on attribute α. The fraction captures the average difference between positive and negative opinions per review.

At this point, we have defined two alternative functions that evaluate the reputation of an item, as captured in its review corpus (see Equations 4 and 5). Let $f(\cdot)$ represent either of these functions. Then, we define the *impact* of a fake review r in the context of a review-corpus R as follows:

$$impact(r, R) = f(R \cup \{r\}) - f(R) \tag{6}$$

The assigned impact value can be positive or negative, depending on whether the fake review was injected to boost or harm the item's reputation. Given our formal definition of impact and authenticity, we can now formalize the *sensitivity* of a review corpus to attacks. Formally, given a corpus R, a stealth threshold λ and a coherence threshold μ, the sensitivity of R to attacks that aim to *boost* the item's reputation is defined as:

$$\begin{aligned} &sensitivity^{(+)}(R, \mu, \lambda) = \\ &\max_{r}\{impact(r, R) : stealth(r, R) \geq \lambda, coh(r) \geq \mu, \} \end{aligned} \tag{7}$$

To compute the sensitivity to attacks that aim to *harm* the item's reputation, it is sufficient to find the review with the *minimum* possible impact, under the given constraints. In our experiments, we use this measure to evaluate the vulnerability of real datasets. Observe that we do not consider readability, since it can be optimized independently.

3 Attack Strategies

Next, we formalize two attack strategies that consider authenticity and impact. For the remainder of this section, we assume attacks that want to *boost* the reputation of a given item. The analysis for attacks that attempt to *hurt* the item trivially follows.

First, we formalize the first attack strategy, which we refer to as the *single-item attack*:

Problem 1. **[Single-Item Attack]:** Given the corpus of reviews R on an item and a reputation function $f(\cdot)$, we want to inject a fake review r^* into R so that $stealth(r^*, R) \geq \lambda$, $coherence(r) \geq \mu$, $readability(r) \geq \nu$ and:

$$impact(r^*, R) = \max_{r}\{impact(r, R)\} \tag{8}$$

If the attacker is interested in short-term results or is confident in the inability of the review management system to identify fake reviews, λ can be set to a low value. Otherwise, a higher value for λ is appropriate. The parameters μ and ν are similarly tuned to determine coherence and readability. In our experiments, we show how these parameters can be learned from real review corpora.

By focusing exclusively on the review corpus of the target, single-item attacks overlook a valuable resource: the reviews on the target's *competitors*. Consider the case when the target item I competes with a competitor I' for the same market share. A single-item attack would try to increase the market share of I by improving its reputation. However, the same can be accomplished by attacking the corpus of I'. By damaging the competitor's reputation, the attack will make it less appealing to customers, who are then likely to turn to I. Formally, we define this attack strategy as follows:

Problem 2. **[Competitor Attack]:** Let $\mathcal{C}$ be the set of immediate competitors for the target item I, and let $\mathcal{R}_{\mathcal{C}}$ be the respective collection of review corpora for the items in $\mathcal{C}$. Then, given a reputation function $f(\cdot)$, we want to inject a fake review r^* into a corpus $R^* \in \mathcal{R}_{\mathcal{C}}$, so that $stealth(r^*, R^*) \geq \lambda$, $coherence(r) \geq \mu$, $readability(r) \geq \nu$ and

$$impact(r^*, R^*) = \min_{r, R \in \mathcal{R}_{\mathcal{C}}}\{impact(r, R)\} \tag{9}$$

We observe that a solution for Problem 1 can be easily extended to optimally solve Problem 2 (i.e. by identifying the review with the highest negative impact for each competitor in $\mathcal{C}$). Thus, we focus on solving Problem 1. We observe that readability is completely independent of impact, and can thus be optimized separately. Thus, we want to find the review with the maximum impact, among those that respect the thresholds λ, μ and ν. As we show next, a single-item attack can be intuitively formed as a 0-1 integer linear programming problem. First, we consider the maximization of the review's impact for each of the two reputation functions considered in our work.

For the rating-based reputation function $f_1(\cdot)$, we want the injected fake review to maximize the average rating over the entire corpus. This trivially translates into submitting the review with the highest possible rating, as allowed by the constraints. For the attribute-based reputation function $f_2(\cdot)$, maximizing $impact(r, R)$ translates into maximizing $f_2(R \cup \{r\})$. Thus, we want to maximize:

$$\propto f_2(R \cup \{r\}) = \sum_{\alpha \in \mathcal{A}} P(\alpha, R) + x_\alpha - N(\alpha, R) - y_\alpha \qquad (10)$$

where:

$$y_\alpha + x_\alpha \leq 1, y_\alpha, x_\alpha \in \{0, 1\} \qquad (11)$$

x_α is set to 1 when the review expresses a positive opinion on α. Similarly, $y_\alpha = 1$ for a negative opinion. The $y_\alpha + x_\alpha \leq 1$ constraint ensures that at most one of the two variables can be set to 1. Now, based on Eq. 1, the stealth constraint can be written as:

$$\sum_{\alpha \in \mathcal{A}} x_\alpha \times [P(\alpha, R) - \lambda(P(\alpha, R) + N(\alpha, R))] + \qquad (12)$$
$$y_\alpha \times [N(\alpha, R) - \lambda(P(\alpha, R) + N(\alpha, R))] \geq 0$$

where x_α and y_α are the variables introduced in Equation 10 above.

Finally, we need to incorporate the coherence threshold μ. Considering Eq. 2, we need to separately consider the assignment of a positive, negative or neutral rating:

$$\text{positive: } \sum_{\alpha \in \mathcal{A}} x_\alpha / (\sum_{\alpha \in \mathcal{A}} x_\alpha + y_\alpha) \geq \mu \quad (13) \qquad \text{negative: } \sum_{\alpha \in \mathcal{A}} y_\alpha / (\sum_{\alpha \in \mathcal{A}} x_\alpha + y_\alpha) \geq \mu \quad (14)$$

$$\text{neutral: } 1 - \left| \sum_{\alpha \in \mathcal{A}} x_\alpha - \sum_{\alpha \in \mathcal{A}} y_\alpha \right| / \left(\sum_{\alpha \in \mathcal{A}} x_\alpha + y_\alpha \right) \geq \mu \quad (15) \qquad (13)$$

If multiple ratings satisfy the coherence threshold μ, then we simply choose the rating that leads to the highest impact.

For both reputation functions, standard techniques for binary integer programming can be used to create the reviews with the highest impact. In order to obtain the solution to Problem 1 for the reputation function $f_2(\cdot)$, we need to solve the optimization problem that maximizes Equation 10, satisfies both constraints given by Equations 11 and 12, and also satisfies *at least one* of three inequalities defined in Equations 13, 14 and 15. For the reputation function $f_1(\cdot)$, we only need to check whether there is an assignment of values for the variables $x_\alpha, y_\alpha, \forall\alpha \in \mathcal{A}$ that satisfies the inequality system consisting of Equations 11, 12 and 13. If no such a solution exists, then we know that there is no positive-rating review that respects the given thresholds. Then, the only option is to lower the thresholds. If a solution exists, we simply assign the maximum possible rating that falls within the "positive" group (e.g. 4 or 5 in the 5-star scale). For this function, we observe that an adversary can bypass the stealth and coherence thresholds by submitting a review that does not include any opinions (just irrelevant or generic text). This can be handled by an additional constraint, asking for at least M opinions to be expressed in a valid review. We explore this direction in our experiments.

4 Experiments

In this section, we present the experiments that we conducted to evaluate our methodology. In our evaluation, we use the openly available data introduced by Jindal and [17]. The dataset consists of more than 5.8 million reviews from Amazon.com. For our evaluation, we focus on the domains of MP3 players and Digital Cameras. These domains were chosen based on the plethora of available reviews. Other than that, our study does not benefit from this choice. In particular, we extracted the reviews for 666 different MP3 players and for 589 cameras. For each review, we retrieve the star rating and the actual text. We refer to these datasets as `MP3` and `CAM`, respectively. The method by Ding et al. [4] was used for opinion extraction. The authors show that their method is superior to previous state-of-the art techniques. Nonetheless, our methodology is compatible with any other alternative method for this task [9, 22].

4.1 Authenticity

In this experiment, we study the observed values of the three components of authenticity: stealth, coherence and readability. This evaluation provides insight on the tuning of the respective thresholds (λ, μ, ν). First, we compute the average value for each measure, over all items in each of the two datasets. We then present the distribution of the three measures across different value intervals in Figure 1. As shown in the figure, the vast majority of the reviews exhibited high readability values (60-80). For comparison,

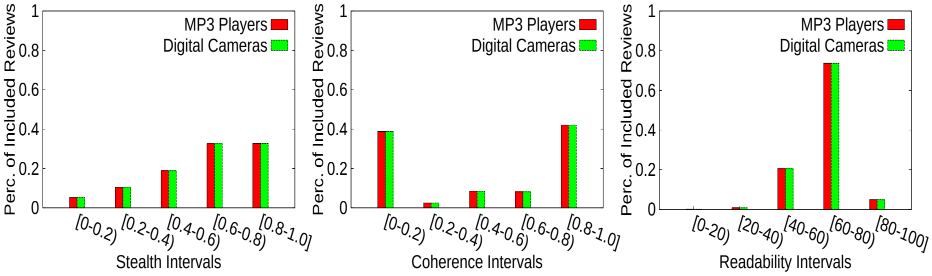

Fig. 1. Distribution of stealth, coherence and readability in real review corpora.

consider that the readability of the popular *Reader's Digest* is about 65, while *Time* magazine scores at about 52 [1]. For coherence, two major clusters emerged from the analysis: about 40% of the reviews achieved high values, indicating that their ratings matched the opinions expressed in their text. However, another 40% of the reviews exhibited very low values (0-0.2). This cluster can serve as the starting point for a review-detection algorithm. However, its large size also hints at the ambiguity of the 5-star scale, which has been pointed out by others in the past [2]. Further, the fact that the ratings assigned by users are often discordant with the expressed opinions, motivates the use of more sophisticated reputation functions than the standard star-scale. With respect to stealth, about 65% of the reviews achieved a value greater than 0.6, while the rest are those that generally contradict the opinions expressed by the majority of the reviewers.

4.2 Corpus Sensitivity

In this experiment we use our sensitivity measure (Eq. 7) to evaluate the vulnerability of real review corpora to fake reviews.

[Attribute-based reputation]: We begin our analysis with the attribute-based reputation function (Eq. 5). For the first experiment, the coherence threshold is set to $\mu = 0.8$ (in accordance with the findings of the previous experiment). Then, for different values of the stealth threshold λ, we compute the sensitivity of each corpus to positive (POS) and negative (NEG) fake reviews. The results are shown in Figure 2, where we show the average sensitivity over all items in each dataset. As can be seen in the figure, the maximum sensitivity to fake positive reviews is observed for $\lambda \leq 0.6$. For higher values the impact declines, as it becomes harder to included positive opinions while respecting the stealth threshold. The respective results on the sensitivity to negative fake reviews are very interesting, with the impact curve declining a lot faster. This can be explained by the bias toward positive reviews that ails most review corpora [12]. In other words, the plethora of positive opinions makes it harder for a negative review to maintain stealth. This finding indicates that, for an attacker who wants to hurt an item's reputation, a competitor-attack that tries to boost the reputation of its competitors can be more effective to an attack that tries to directly hurt the item's reputation.

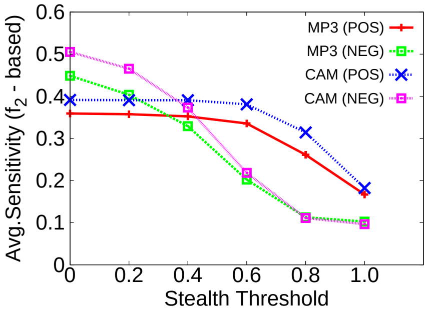

Fig. 2. Sensitivity Vs. Stealth

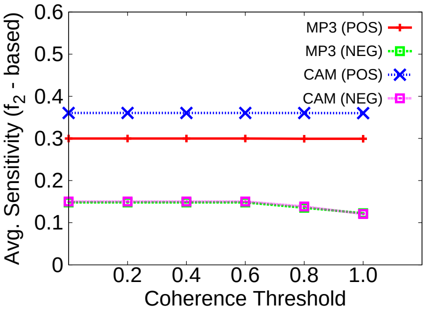

Fig. 3. Sensitivity Vs. Coherence

To complete our evaluation of sensitivity, we fix the stealth threshold to $\lambda = 0.6$ and tune the coherence threshold μ. The results are shown in Figure 3. The figure illustrates that the coherence threshold has little or no effect on the sensitivity of the corpus. In other words, it is typically feasible for an adversary to create fake (positive or negative) review that abides by even the strictest of thresholds.

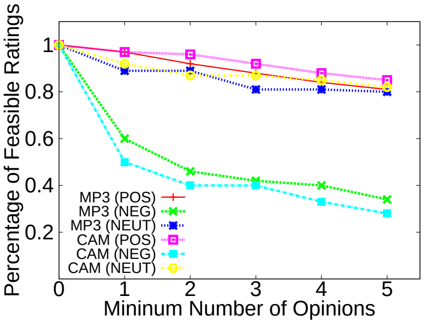

Fig. 4. Rating feasibility Vs. Min Number of Opinions in a fake review (M)

	Single-Item	Competitor
Success Perc.	71%	91%
#Reviews (Avg.)	35.8	15.8
#Reviews (Stdev.)	36.1	12.8

Fig. 5. Results for the two Attack Strategies

[Rating-based reputation]: As discussed in our analysis on the rating-based function, it is sufficient to check which ratings are feasible under the stealth and coherence thresholds. However, an adversary can bypass the stealth and coherence thresholds by submitting a review that does not include any opinions. As mentioned in our analysis, this can be handled by asking for at least M opinions to be expressed in a valid review. Next, we explore the effect of this parameter on the sensitivity of the corpus. As before, the stealth and coherence threshold as set to $\lambda = 0.6$ and $\mu = 0.8$, respectively. In Figure 4, we report the percentage of corpora for which a positive, neutral and negative rating was feasible, for different values of M (i.e. the minimum number of required opinions).

As can be seen by the figure, M has a minor effect on the feasibility of positive and neutral reviews. On the other hand, the feasibility of negative ratings clearly decreases as M rises. This can be attributed to the bias toward positive evaluations in real review corpora, which also emerged for the attribute-based function. For a negative rating to be possible as more opinions are required, the coherence threshold requires the majority of these opinions to be negative. However, the positive bias makes it harder for a review with numerous negative opinions to respect the stealth threshold.

4.3 Comparing Attack Strategies

Finally, we compare the two attack strategies that we formalized in our work (*single-item* and *competitor*). For this experiment, we use all the corpora from MP3. For every possible pair, we identify the item with the lowest reputation, based on the attribute-based function $f_2()$ (the findings for the rating-based function were similar and are omitted for lack of space). We refer to this item as the *client*, while the other item of the pair is referred to as the *victim*. Our goal is to inject enough fake reviews for the *client* to achieve an equal or higher score than the *victim*. Using the *single-item attack*, we repeatedly inject the fake review with the highest positive impact until the goal is reached. In addition to this option, the *competitor attack*, allows us to inject reviews with the highest possible negative impact to the *victim*'s corpus. For this strategy, we greedily choose the option that brings the scores of the two items closer. We record the number of fake reviews required for each strategy. If more than a 100 reviews are required to reach the goal, the attack is considered a failure. The results are shown in Figure 5. The first observation is that the competitor-attack reaches the goal for 90% of the pairs, cleary outperforming the single-item strategy. For those successful pairs, the two methods required about the same number of fake reviews on average. Finally, the high standard variation illustrates that sensitivity can vary across review corpora, a finding that is in accordance with our own previous evaluation.

5 Related Work

A significant amount of work has recently been devoted to the identification of fake reviews [17, 11, 10]. In addition to the textual and linguistic features of reviews, relevant papers have explored different aspects of the topic, including review-ratings [26, 3], group spamming [18], and atypical (suspicious) reviews [13, 15]. Atypical reviews are an interesting concept that we also consider in our own work. Another relevant line of work explores the behavior of reviewers [16, 25]. By tracking a user's reviewing and rating patterns, these methods aim at finding suspicious behavior that hints to malice or bias against particular brands or products. Our contribution is complementary to that made by these papers: while this line of work examines the problem from the defensive point of view (trying to identify and eliminate spam or malicious reviews) our work is the first to explore the opposite perspective, examining the creation of fake reviews from the attacker's side. Finally, our work has ties to the extensive literature on opinion mining and sentiment analysis [20]. In particular, we employ the method by Ding et al. [4] for the extraction of opinions from reviews. Nonetheless, our framework is compatible with any alternative technique for opinion extraction [9, 22].

6 Conclusion

In this paper we presented the first study of fake reviews from the perspective of the attacker. We formalized the factors that determine the success of an attack and explored different attack strategies. Our analysis and experimental findings on real datasets provide valuable insight on the domain of fake reviews, and can be used as the basis for the immunization of review corpora and the improvement of fraud-detection techniques.

References

1. http://en.wikipedia.org/wiki/flesch_kincaid_readability_test.
2. http://www.wellesleywinepress.com/2011/09/on-5-star-system-and-100-point-scale.html.
3. C. Dellarocas. Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. EC '00, pages 150–157, New York, NY, USA, 2000. ACM.
4. X. Ding, B. Liu, and P. S. Yu. A holistic lexicon-based approach to opinion mining. WSDM '08, pages 231–240, New York, NY, USA, 2008. ACM.

5. W. Duan, B. Gu, and A. B. Whinston. Do Online Reviews Matter? - An Empirical Investigation of Panel Data. *Social Science Research Network Working Paper Series*, Nov. 2004.
6. R. Flesch. A new readability yardstick. In *J Appl Psychol 32*, pages 221–224, 1948.
7. R. Flesch. *The Art of Readable Writing*. Harper and Row, 1949.
8. L. Hankin. *The effects of user reviews on online purchasing behavior across multiple product categories*. PhD thesis, 2007.
9. M. Hu and B. Liu. Mining and summarizing customer reviews. KDD '04, pages 168–177, New York, NY, USA, 2004. ACM.
10. N. Jindal and B. Liu. Analyzing and detecting review spam. In *ICDM*, pages 547–552, 2007.
11. N. Jindal and B. Liu. Review spam detection. WWW '07, pages 1189–1190, New York, NY, USA, 2007. ACM.
12. N. Jindal and B. Liu. Opinion spam and analysis. WSDM '08, pages 219–230, New York, NY, USA, 2008. ACM.
13. N. Jindal, B. Liu, and E.-P. Lim. Finding unusual review patterns using unexpected rules. CIKM '10, pages 1549–1552, New York, NY, USA, 2010. ACM.
14. N. Korfiatis, D. Rodrguez, and M.-A. Sicilia. The impact of readability on the usefulness of online product reviews: A case study on an online bookstore. In *Emerging Technologies and Information Systems for the Knowledge Society*, volume 5288, pages 423–432. 2008.
15. T. Lappas and D. Gunopulos. Efficient confident search in large review corpora. In *ECML/PKDD (2)*, pages 195–210, 2010.
16. E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw. Detecting product review spammers using rating behaviors. CIKM '10, pages 939–948, NY, USA, 2010. ACM.
17. J. Mackiewicz. Reviewer motivations, bias, and credibility in online reviews. In S. Kelsey and K. S. Amant, editors, *Handbook of Research on Computer Mediated Communication*, pages 252–266. IGI Global, 2008.
18. A. Mukherjee, B. Liu, J. Wang, N. S. Glance, and N. Jindal. Detecting group review spam. In *WWW (Companion Volume)*, pages 93–94, 2011.
19. M. P. O'Mahony and B. Smyth. Using readability tests to predict helpful product reviews. RIAO '10, pages 164–167, Paris, France, France, 2010.
20. B. Pang and L. Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2:1–135, January 2008.
21. D.-H. Park, J. Lee, and I. Han. The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement. *Int. J. Electron. Commerce*, 11(4):125–148, July 2007.
22. A.-M. Popescu and O. Etzioni. Extracting product features and opinions from reviews. HLT '05, pages 339–346, Stroudsburg, PA, USA. Association for Computational Linguistics.
23. D. A. Reinstein and C. M. Snyder. The influence of expert reviews on consumer demand for experience goods: A case study of movie critics. *Journal of Industrial Economics*, 53(1):27–51, 03 2005.
24. P. Tsaparas, A. Ntoulas, and E. Terzi. Selecting a comprehensive set of reviews. In *KDD*, pages 168–176, 2011.
25. G. Wang, S. Xie, B. Liu, and P. S. Yu. Review graph based online store review spammer detection. *Data Mining, IEEE International Conference on*, 0:1242–1247, 2011.
26. G. Wu, D. Greene, B. Smyth, and P. Cunningham. Distortion as a validation criterion in the identification of suspicious reviews. Technical report, 2010.
27. L. Zhuang, F. Jing, and X.-Y. Zhu. Movie review mining and summarization. CIKM '06, pages 43–50, New York, NY, USA, 2006. ACM.