Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)

Behavioral Analysis of Review Fraud: Linking Malicious Crowdsourcing to Amazon and Beyond

Parisa Kaghazgaran, James Caverlee, Majid Alfifi
Texas A&M University
College Station, TX 77843
{kaghazgaran,caverlee,alfifi}@tamu.edu

Abstract

We exploit the prevalence of malicious review writers on crowdsourcing platforms like RapidWorkers to identify actual fraud reviews on Amazon. Complementary to previous efforts which often rely on proxies for fraud reviews, we present a long-term study of actual fraudulent behaviors in online review manipulation. We find that these malicious reviewers – though often providing seemingly legitimate opinions – do exhibit significant differences from normal reviewers in terms of ratings distribution, length of the reviews, and the burstiness of the reviews themselves. We additionally study the evolution of these reviews, and find striking temporal changes that could support future discovery of these reviewers who may be "hiding in plain sight".

Introduction

User reviews are a cornerstone of how we make decisions. From deciding what movies to view, products to purchase, restaurants to patronize, and even doctors to visit, user review aggregators like Amazon, Netflix, and Yelp shape our experiences. And yet, these reviews are vulnerable to manipulation (Weise 2016). This manipulation threatens to degrade trust in these online platforms and in their products and services. Indeed, many previous efforts have explored methods to uncover this manipulation, often by applying machine learning or graph-based algorithms, e.g., (Prakash et al. 2010), (Wang et al. 2011), (Akoglu, Chandy, and Faloutsos 2013), (Shah et al. 2014), (Jiang et al. 2014), (Ye and Akoglu 2015). These methods typically are built and validated over a dataset of "known" manipulated reviews. And yet, most make one of several critical assumptions:

- *Manual labeling of fake reviews:* In the first approach, judges – often either researchers themselves or a team of labelers at a review site – assess individual reviews to determine if they are fake or not (Mukherjee et al. 2013), (Rayana and Akoglu 2015). These methods sometimes rely on unsupervised algorithms (e.g., the output of a proprietary company algorithm) or on manual and possibly error-prone labeling of fake reviews without access to a ground truth.
- *Ex post analysis of outliers:* A second approach is to validate detection algorithms through ex post analysis of suspicious reviews. Typically, an algorithm is run over a collection of reviews and the top-ranked results are examined (Wang et al. 2011), (Akoglu, Chandy, and Faloutsos 2013), (Ye and Akoglu 2015). This approach tends to focus on highly-visible fake behaviors (e.g., a reviewer who posts dozens of reviews in a period of minutes), but may miss more subtle behaviors.
- *Simulation of bad behavior:* A recent third approach is to *simulate* the behaviors of malicious workers (Ott, Cardie, and Hancock 2013). In this approach, volunteers are asked to imagine themselves as fake review writers and then post fake reviews. While encouraging, this method necessarily lacks insight into the strategies and motivations of actual fake review writers.

We seek to complement these foundational studies by investigating the strategies of a collection of *actual fake review writers*. By monitoring low moderation crowdsourcing sites like RapidWorkers, ShortTask, and Microworkers, we can gain access to a pool of crowd workers who we know for certain have engaged in fake review writing. By examining the behaviors of these fake review writers, we aim to provide the first long-term study of actual fraudulent behaviors in online review manipulation[1]. As an example, consider the following task posted to RapidWorkers:

> **What is expected from workers?**
> Read the product description before writing down a review.
> Go to https://goo.gl/7QfW0h.
> Leave a relevant 5-star review with at least 40 words.
> Provide proof that you left the review yourself.

By monitoring such requests, we can begin to study the behaviors of fraudulent review writers. Although not representative of all types of manipulation, this approach does provide the tantalizing opportunity to study malicious behaviors in the wild.

Through our initial investigation of 100 targeted products

[1] A previous work (Fayazi et al. 2015) has examined a collection of reviews launched from crowdsourcing sites as a snapshot, but without considering the evolutionary behavior of the reviewers.

★★★★★ **Best skin lightening Cream !!!**

This is the best skin lightening cream by Diva for those who want to moisturize and nourish your skin. This cream has green tea extract that lifts the skin. I used the product and could see the wrinkles fading in two weeks. Now my skin looks more lighter now. It is more radiant and vibrate.

Figure 1: An example review written by a crowd worker.

on Amazon, 5,200 reviewers, and 350,000 reviews, we find striking behaviors of malicious reviewers. In many cases, though often providing seemingly legitimate opinions, these reviewers do exhibit significant differences from normal reviewers in terms of ratings distribution, length of the reviews, and the burstiness of the reviews themselves. We additionally study the evolution of these reviews, and find striking temporal changes that could support future discovery of these reviewers who may be "hiding in plain sight".

Collecting Malicious Reviews

We focus in this paper on tasks posted to a single crowdsourcing site – RapidWorkers – that target Amazon. Note that there are many such sites and many additional targets (e.g., Yelp, App Store, Play Store, etc.) (Wang et al. 2012), (Lee, Webb, and Ge 2014). Typically, tasks on these sites pay workers from $0.10 to $1.50 per task, where a single target (e.g., a product on Amazon) may be subject to dozens of fake reviews launched from these crowdsourcing sites. As an example of the type of review that is created, Figure 1 shows a sample of a crowdsourced review for a "skin lightening cream" product sold by Amazon.

Concretely, we crawl all tasks related to promoting products in Amazon from the RapidWorkers platform from July 2016 to November 2016. In total, we identify 100 unique Amazon product IDs. For each of these IDs, we crawl the corresponding product page at Amazon, plus all reviews associated with the product. For each reviewer we encounter, we additionally collect all of their reviews (which may include products beyond those targeted by these crowdsourcing sites). Ultimately, our dataset contains the following information: product ID, review ID, reviewer ID, review title, review content, rating, time-stamp and "verified purchase" flag. In total, we identify 5,200 unique reviewers and 350,000 unique reviews.

Fraudulent vs Non-Fraudulent Reviewers

Our dataset naturally contains a mix of reviewers and their reviews: some are legitimate reviews, some are the result of targeted crowdsourced efforts, while others may also be fraudulent but outside the purview of our sampling method (e.g., launched via an unobservable channel like private email). Hence, we make a conservative assumption for the rest of this paper: We consider a reviewer to be a *fraudulent reviewer* if they have reviewed **two or more products** that have been targeted by a crowdsourcing effort. Intuitively, workers may aim to maximize their income by participating in many tasks (and hence, targeting many products). On the other hand, it is unlikely that a random user will write a legitimate review on two different crowdsourcing products in a short period of time, considering Amazon's selection of millions of products (Grey 2015).

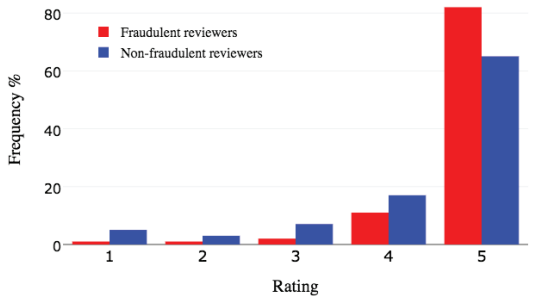

Figure 2: Ratings distribution: fraudulent reviewers tend to give more 4 and 5-star reviews.

Making this conservative assumption, we identify 625 of the 5,200 reviewers as *fraudulent reviewers*. Of the remaining 4,575 reviewers, there is certainly a mix of fraudulent and legitimate reviewers. We make a further assumption that any reviewer who has reviewed only one of the crowdsourcing products and has actually purchased the product itself (via the "verified purchase" attribute in our dataset) is a *non-fraudulent reviewer*. Of course, there may still be some unknown fraudulent reviewers in this set of 2,800 reviewers, but it gives us a baseline to compare against the clearly prolific fraudulent reviewers.

Observations

Given these two sets of reviewers – fraudulent and non-fraudulent – how do they behave? Since these two groups have reviewed a similar cohort of products, we expect that differences in behavior are attributable mainly to these differing motivations, and not due to differences in products themselves (e.g., a health and beauty product may attract a different reviewer profile from a home improvement product). In this section we investigate differences due to ratings, the characteristics of the reviews themselves, as well an examination of the evolution of these reviews.

Ratings. We begin with Figure 2, which shows the ratings distribution for reviews written by our two types of reviewers. Echoing previous studies e.g., (Hooi et al. 2015) we see that crowdsourcing workers tend to write 4 or 5-star reviews. While crowdsourcing efforts could be targeted at suppressing the ratings for a competitor, we see instead that most efforts focus on promotion. Compared to legitimate reviewers, the rate of 5-star reviews is 17% higher for fraudulent reviewers.

Review Characteristics. We next turn to the characteristics of the reviews – in terms of the length of the review, the burstiness of reviews written by a reviewer, the fraction of reviews that have been for products actually purchased, as well as a preliminary investigation into their content.

Review Length. We see in Figure 3 the distribution of the review length in terms of number of characters between the

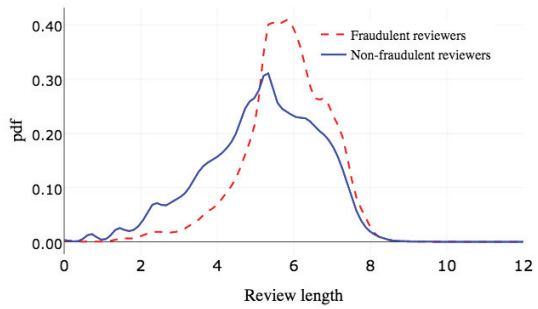

Figure 3: Review length distribution (log-scale) in terms of number of characters per review.

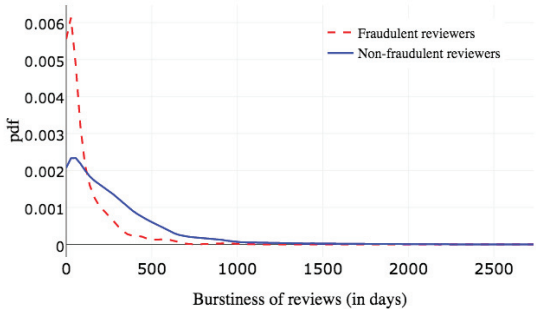

Figure 4: Burstiness of reviews: the x-axis measures the timestamp standard deviation per reviewer. The distribution for fraudulent reviewers is left-skewed indicating that their reviews tend to be posted in bursts.

two groups (note the graph is on a log-scale). Contrary to our intuition that reviews by fraudulent reviewers would be relatively short, we see that these reviews are relatively lengthy. We can attribute this to task requestors requiring a minimum character length for payment.

Burstiness of Reviews. Intuitively, crowd workers may seek to complete several tasks in a short time to maximize their payoff. Hence, for each reviewer we measure the standard deviation of the timestamp for that person's reviews. We plot the distribution for this "burstiness", as seen in Figure 4. In this case, a small standard deviation corresponds to many reviews being posted in a short time window, whereas a higher standard deviation corresponds to reviews posted over a long time period (and hence, lacking burstiness). We can see that the distribution for fraudulent reviewers is left-skewed indicating that their reviews tend to be posted in bursts.

Actual Purchases? We also observe differences in amount of "verified" purchases between two groups. Recall that our assignment of reviewers to the non-fraudulent category included the requirement that the review on a crowdsourced targeted product had been verified; here we consider the verified status of all additional products reviewed. As we see in Figure 5, most reviews by fraudulent reviewers are for products that do not have an associated "verified" purchase. Nearly 7-times as many legitimate reviewers provide

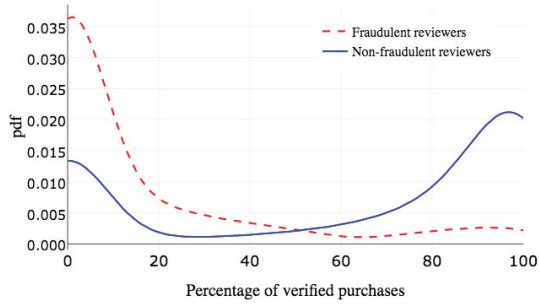

Figure 5: Verified purchases: Most reviews by fraudulent reviewers are for products that do not have an associated "verified" purchase.

reviews on 100% of verified purchases compared to fraudulent reviewers.

Conveying Personal Experiences. As a first look at the content of the reviews themselves, we calculate the fraction of reviews that contain a first-person pronoun. The intuition is that crowdsourcing workers may try to appear normal by referring to their own experiences with a product through the use of personal pronouns like *I, my, mine, we, ours,* and *us*. In this way, the crowd workers can convey the impression that they have actually used the product. To confirm this intuition, we calculate the frequency of these pronouns in both groups of reviews: we find that 66% of reviews by fraudulent reviewers contain first-person, compared with only 42% of reviews by legitimate reviewers.

Evolution of Review Behavior. Given these initial insights into the differences between fraudulent and non-fraudulent reviewers, we turn in this section to an examination of the lifecycle of these reviewers. Do they mimic themselves? Do they demonstrate signs of significant changes overtime?

Self-similarity. We first measure how much a reviewer's language mimics previous reviews they have written. Perhaps fraudulent reviewers write according to a simple "template", and so new reviews tend to repeat language used in previous ones? Here, we measure the lexical overlap between each review and the previous 10 reviews written by the same reviewer using the Jaccard similarity (JS). Each 10 sequential reviews form one life stage. E.g.,

$$JS = \frac{\sum_{i=1}^{10} \frac{|r_i \cap r|}{|r_i \cup r|}}{10}$$

defines the self-similarity for review r in the first life stage of its author. Figure 6 shows that non-fraudulent tend not to repeat themselves (the left-skewed curve); whereas fraudulent reviewers tend to rely on repeated keywords or phrases. Intuitively, reviewers engaged in crowd-launched manipulation tend to mimic themselves over time since they are not actually experienced with actual product.

Linguistic evolution. Finally, we explore how the reviewers linguistically evolve over time. Using reviews from January 2015 to November 2016, we build a bigram language model for each month (Danescu-Niculescu-Mizil et al. 2013) that represents the overall background language used for a partic-

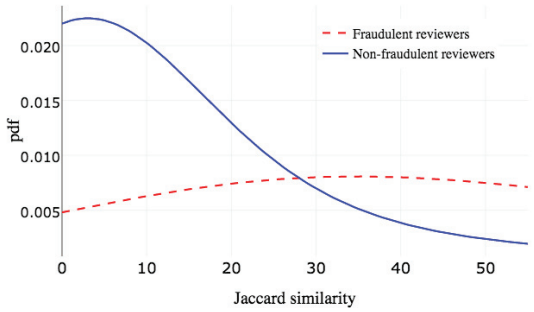

Figure 6: Self-similarity distribution for reviewers with at least 50 reviews.

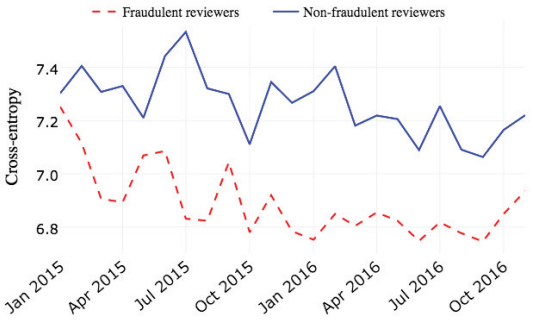

Figure 7: Evolution of Language

ular period. Then we can quantify how the language of a single review (r) differs from the background language model (BLM) of the month (m) it was written by calculating the cross entropy between r and BLM_m:

$$H(r, BLM_m) = -\frac{1}{N} \sum \log P_{BLM_m}(b_i)$$

where b_i are the bigrams of r and $P_{BLM_m}(b_i)$ is the probability of b_i based on that month's language model. Cross entropy captures how surprising a review r is with respect to the language used by the rest of the reviewing community: higher values indicate a review differs more. Figure 7 shows how both fraudulent and non-fraudulent reviewers begin in January 2015 with nearly the same degree of difference from the background language model. Over time, we see that the non-fraudulent reviewers are fairly consistent, whereas the fraudulent reviewers begin to deviate more and more (with a lower cross entropy). Echoing the result above for self-similarity, we surmise that fraudulent reviewers begin to reuse common linguistic patterns, whereas non-fraudulent reviewers continue to innovate linguistically through new experiences with new products.

Conclusion and Future Work

We have explored how monitoring tasks on sites like RapidWorkers can uncover fraudulent reviewers on sites like Amazon. This framework complements previous efforts by providing a new approach for identifying these types of reviewers. Our behavioral analysis of these actual fake review writers has also uncovered clues that may aid in their detection. In our ongoing work, we are expanding our coverage both in terms of crowdsourcing sites and targets of manipulation (e.g., App Store, Play Store, Yelp). We are also eager to further explore how linguistic evolution may provide new insights into the strategies of review manipulation.

Acknowledgement. This work was supported in part by AFOSR grant FA9550-15-1-0149.

References

Akoglu, L.; Chandy, R.; and Faloutsos, C. 2013. Opinion fraud detection in online reviews by network effects. In *ICWSM*.

Danescu-Niculescu-Mizil, C.; West, R.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *WWW*.

Fayazi, A.; Lee, K.; Caverlee, J.; and Squicciarini, A. 2015. Uncovering crowdsourced manipulation of online reviews. In *SIGIR*.

Grey, P. 2015. How many products does amazon sell?, https://export-x.com, last access: 01/10/2017.

Hooi, B.; Shah, N.; Beutel, A.; Gunneman, S.; Akoglu, L.; Kumar, M.; Makhija, D.; and Faloutsos, C. 2015. Birdnest: Bayesian inference for ratings-fraud detection. *arXiv*.

Jiang, M.; Cui, P.; Beutel, A.; Faloutsos, C.; and Yang, S. 2014. Inferring strange behavior from connectivity pattern in social networks. In *PAKDD*.

Lee, K.; Webb, S.; and Ge, H. 2014. The dark side of micro-task marketplaces: Characterizing fiverr and automatically detecting crowdturfing. In *ICWSM*.

Mukherjee, A.; Venkataraman, V.; Liu, B.; and Glance, N. S. 2013. What yelp fake review filter might be doing? In *ICWSM*.

Ott, M.; Cardie, C.; and Hancock, J. T. 2013. Negative deceptive opinion spam. In *HLT-NAACL*.

Prakash, B. A.; Sridharan, A.; Seshadri, M.; Machiraju, S.; and Faloutsos, C. 2010. Eigenspokes: Surprising patterns and scalable community chipping in large graphs. In *PAKDD*.

Rayana, S., and Akoglu, L. 2015. Collective opinion spam detection: Bridging review networks and metadata. In *SIGKDD*.

Shah, N.; Beutel, A.; Gallagher, B.; and Faloutsos, C. 2014. Spotting suspicious link behavior with fbox: An adversarial perspective. In *ICDM*.

Wang, G.; Xie, S.; Liu, B.; and Philip, S. Y. 2011. Review graph based online store review spammer detection. In *ICDM*.

Wang, G.; Wilson, C.; Zhao, X.; Zhu, Y.; Mohanlal, M.; Zheng, H.; and Zhao, B. Y. 2012. Serf and turf: crowdturfing for fun and profit. In *WWW*.

Weise, E. 2016. Amazon bans 'incentivized' reviews, goo.gl/k8woqd, last access: 01/10/2017.

Ye, J., and Akoglu, L. 2015. Discovering opinion spammer groups by network footprints. In *ECML-PKDD*.