Noname manuscript No.
(will be inserted by the editor)

Characterizing the Dynamics and Evolution of Incentivized Online Reviews on Amazon

Soheil Jamshidi · Reza Rejaie · Jun Li

Received: date / Accepted: date

Abstract During the past few years, sellers have increasingly offered discounted or free products to selected reviewers of e-commerce platforms in exchange for their reviews. Such incentivized (and often very positive) reviews can improve the rating of a product which in turn sways other users' opinions about the product. Despite their importance, the prevalence, characteristics, and the influence of incentivized reviews in a major e-commerce platform have not been systematically and quantitatively studied.

This paper examines the problem of detecting and characterizing incentivized reviews in two primary categories of Amazon products. We describe a new method to identify Explicitly Incentivized Reviews (EIRs) and then collect a few datasets to capture an extensive collection of EIRs along with their associated products and reviewers. We show that the key features of EIRs and normal reviews exhibit different characteristics. Furthermore, we illustrate how the prevalence of EIRs has evolved and been affected by Amazon's ban. Our examination of the temporal pattern of submitted reviews for sample products reveals promotional campaigns by the corresponding sellers and their effectiveness in attracting other users. We also demonstrate that a classifier that is trained by EIRs (without explicit keywords) and normal reviews can accurately detect other EIRs as well as implicitly incentivized reviews. Finally, we explore the current state of explicit reviews on Amazon. Overall, this analysis sheds insightful light on the impact of EIRs on Amazon products and users.

Keywords Incentivized Online Reviews, Machine Learning, Modeling, Amazon, Online Review

Department of Computer and Information Science
University of Oregon
E-mail: {Jamshidi, Reza, Lijun}@cs.uoregon.edu

1 Introduction

As the popularity of online shopping has rapidly grown during the past decade, the shoppers have increasingly relied on online reviews and rating provided by other users to make more informed purchases. In response to shoppers' behavior, product sellers have deployed various strategies to attract more positive reviews for their products as this could directly affect the popularity of these products among users and thus their ability to sell more products online. Several prior studies have examined different aspects of online reviews including fake or spam [10,17,11,20,16,2] and also biased and paid reviews [25,27,28,21,6] in different online shopping platforms.

The importance of online reviews has also prompted major e-commerce sites (*e.g.*, Amazon) to implement certain policies to ensure that the provided user reviews and ratings are legitimate and unbiased to maintain the trust of online shoppers. In response to these policies, seller's strategies for boosting their product rating have further evolved. In particular, in the past few years, some sellers have increasingly offered discounted or free products to selected online shoppers in exchange for their (presumably positive) reviews. Such reviews are called *incentivized reviews* [21,23]. Major e-commerce sites such as Amazon require reviewers to disclose any financial or close personal connection to the brand or the seller of the reviewed products [3]. However, it is unlikely that average shoppers who solely rely on product ratings notice the biased nature of such reviews. Intuitively, the reviewers who provide incentivized reviews may behave differently than other reviewers for the following reasons: *(i)* they might feel obligated to post positive reviews as the products are provided for free or with a considerable discount, *(ii)* their expectations might be lower than other users as they do not pay the full price, and *(iii)* they do not often consider the long-term usage of the product (*e.g.*, product

return or customer service) in their reviews. The presence of such incentivized reviews in Amazon has been reported in 2016 [23], however, to our knowledge, the prevalence of incentivized reviews, their characteristics, and their impact on the ecosystem of a major e-commerce site have not been systematically and quantitatively studied. Although Amazon has officially banned submission of incentivized reviews in October of 2016 [1], it is important to study such reviews to be able to determine whether Amazon's new policy solved the issue or just forced reviewers to go undercover. To tackle this important problem, this paper focuses on capturing and characterizing several aspects of incentivized reviews in the Amazon.com environment. We leverage the hierarchical organization of Amazon products into categories and subcategories and collect all the information for top-20 best-seller products in all subcategories of two major categories. We then present a method to identify explicitly incentivized reviews (EIRs) on Amazon by identifying a number of textual patterns that indicate explicitly incentivized reviews. We carefully capture and fine-tune these textual patterns using a regular expression. We then use these patterns to identify a large number of EIRs along with their associated products and reviewers. We characterize the key features of EIRs and associated reviewers and products.

Our analysis demonstrates the effect of Amazon ban on the prevalence of EIRs as well as the difference between the features of EIRs and normal reviews. We also examine the temporal pattern of EIR, and non-EIR reviews that a product receives and a reviewer produces to address two questions: *(i)* how the arrival pattern of EIRs for a specific product affects the level of interest (*i.e.*, rate of non-EIRs and their assigned rating) among other users, and *(ii)* how individual reviewers over time become engaged in providing EIRs. Given the apparent gap between the features of normal reviews and EIRs, we examine whether machine learning techniques can detect these differences to identify both explicitly or implicitly incentivized reviews. We show that such a technique can indeed detect other explicitly and implicitly incentivized reviews. Additionally, we investigate the current status of identified EIRs and the ability of sellers to solicit incentivized reviews in today's Amazon platform.

This journal paper supersedes the earlier version of this study that appeared in ASONAM'18 [9] and incorporates the following extensions: *(i)* Adding a few new characterizations including the self-similarity of review content, the timeline of review submissions and the association of EIR reviewers and sellers; *(ii)* Adding other samples of review submission patterns by EIR reviewers, *(iii)* Revising and improving the machine learning model for predicting EIR reviews, and more rigorously evaluating the model, including feature importance and classification confidence; *iv* Investigating the current status of inferred EIRs on Amazon and unveiling new seller strategies to recruit incentivized reviews on Amazon.

The rest of this paper is organized as follows: We describe our data collection technique and our datasets in Section 2. Section 3 presents our method for detecting EIRs. We characterize several aspects of EIRs and their associated products and reviews in Section 4. Section 5 discusses the temporal patterns of EIRs and non-EIRs that are submitted for individual products or produced by individual reviewers. Section 6 presents our effort for automated detection of other explicitly or implicitly incentivized reviews using machine learning techniques. The current state of EIR reviews is explored in Section 7. We present a summary of the most relevant prior work and how they differ from this study in Section 8. Finally, Section 9 concludes the paper and summarizes our future plans.

2 Data Collection and Datasets

This section summarizes some of the key challenges with data collection and then describes our methodology for collecting representative datasets that we capture and use for our analysis. Amazon web site organizes different products into categories that are further divided into smaller subcategories. Each product is associated with a specific seller. A user who writes one (or multiple) review(s) for any product is considered a reviewer of that product. For each entity (*i.e.*, user, review or product), we crawled all the available attributes on Amazon as follows:

- Reviews' attributes: review id, reviewer id, product id, Amazon Verified Purchase (AVP) tag, date, rating, helpful votes, title, text, and link to images.
- Products attributes: product id, seller id, price, category, rating, and title.
- Reviewers' attributes: reviewer id, rank, total helpful votes, and publicly available profile information.

In particular, *AVP tag* of a review indicates whether the corresponding reviewer has purchased this product through Amazon and without deep discount or not [4].

There are a few challenges for proper collection and parsing of this information from Amazon. First, there is a very large number of product categories where the format, available fields for products, and tendency of users to offer reviews widely vary across different categories. Furthermore, we need to comply with the ethical guidelines as well as the enforced rate limits by Amazon servers for crawlers which makes it impossible to collect the reviews for all products within a reasonable window of time. To cope with these challenges, we collect three datasets where each one provides representative samples of products, reviews and reviewers in multiple rounds since March 2015.

Table 1 Basic Features of Our Datasets

	Products (DS1)	EIRs (DS2)	Normal Reviews	Reviewers (DS3)
Reviews	3,797,575	100,086	100,086	216,545
Reviewers	2,654,048	39,886	98,809	2,627
Products	8,383	1,850	1,641	184,124

Sample Products (DS1): We focus on two popular categories of products, namely *Electronics* and *Health & Personal Care* since they have a large number of sub-categories and products that receive many reviews. To make the data collection manageable and given the skewed distribution of reviews across products, we only capture all the information for the top-20 [1]best seller products in each sub-category in the above two categories from *Amazon.com*. While these products represent a small fraction of all products in these two categories, the top-20 products receive most of the attention (#reviews) from users and enable us to study incentivized reviews. We refer to this product-centric dataset as *DS1*.

Sample EIRs (DS2): Using our technique for detecting Explicitly Incentivized Reviews (EIR) that is described in Section 3, we examine all the reviews associated with products in DS1 and identify any EIRs among them. We refer to this set of EIRs as DS2 dataset.

Normal Reviews: After excluding EIRs, we examine the remaining reviews for products in DS1 and consider each review as normal if it is not among EIRs and *(i)* associated with an Amazon Verified Purchase, *(ii)* submitted on the same set of products that received EIRs, and *(iii)* submitted by users who have not submitted any EIRs. We rely on this rather conservative definition of normal reviews to ensure that they are clearly not incentivized. We identified 1,214,893 normal reviews and then selected a random subset of them (the same number as EIRs). We refer to these selected reviews as our normal review dataset that serves as the baseline for comparison with EIRs in some of our analysis.

Incentivized Reviewers (DS3): To get a complete view of sample incentivized reviewers, we randomly select 10% of reviewers associated with the reviews in *DS2* dataset. For each selected reviewer with a public profile, we collect their profile information and all of their available reviews. Overall, we collect this information for 2,627 reviewers (6.59% of them) and only consider their reviews for our analysis. The review system in Amazon is very dynamic since products and users become deactivated or unavailable over time, and reviews might be removed if they have low quality or violate the guidelines. Furthermore, Amazon also changes its hashing mechanism which results in inconsistent identifiers for different items (*e.g.*, reviews) overtime. Because of these dynamics, we observe a very small overlap between products (48.43%) and reviews (0.23%) of DS3 that are also present in DS2. The DS1, DS2, and Normal review datasets are collected in December 2016 while DS3 was collected later in January 2018.

[1] https://www.amazon.com/gp/bestsellers/

3 Detecting Explicit Incentivized Reviews

Automated identification (or labeling) of incentivized reviews requires a reliable indicator in such reviews. To this end, we first focus on reviews in which the reviewer *explicitly* indicates his/her intention for writing the review in exchange for a free or discounted product. Such an indication must be provided in the reviews since Amazon requires that reviewers disclose any incentive they might have received from the sellers [3]. Furthermore, these reviewers also include such incentives in their reviews to attract more sellers to offer them similar incentives in exchange for their reviews to promote their products. Our manual inspection[2] of a large number of reviews revealed that many reviewers indeed explicitly state their incentive for writing their reviews. These reviews contain some variants of the following statements: *"I received this product at a discount in exchange for my honest/unbiased review/feedback."* To capture all variants of such statements, we select any review that matches the following regular expression in a single sentence of the review:

$$'(sent|receive|provide)[\hat{}\backslash.!?] * (discount|free|in-trade|in-exchange)[\hat{}\backslash.!?] * (unbiased|honest)[\hat{}\backslash.!?] * (review|opinion|feedback|experience)'$$

Among all the 3.79M reviews in the DS1 dataset, 100,086 reviews submitted by 39,886 users on 1,850 products match some variants of the above regular expression in one sentence. We consider these 100,086 reviews as EIRs and group them in our DS2 dataset.

We also considered a more relaxed setting where reviews could have the above regular expression across multiple sentences. This strategy tags 325,043 reviews from 210,198 users on 7,059 products as EIR. However, our careful inspection of many of the newly-identified EIRs by this more flexible strategy revealed that some of them are non-incentivized reviews that happen to match the regular expression. To avoid

[2] Our manual inspection process was conducted in multiple rounds as follows: We first select all the reviews that contain our target keywords (*e.g.*, *free*, *discount*) to create a pool. Then, we select 100 random samples of reviews from this pool to manually inspect in each round. As EIRs tend to contain some variants of the same disclaimer sentence, our manual inspection quickly identifies such signatures, and use them to automatically identify reviews in the pool that contain similar signatures. The examination of these reviews also reveals false alarm cases.

any such false-positives in our EIRs, we adopt a conservative strategy and only consider a review as EIR if the desired pattern is detected within a single sentence.

EIR-Aware Reviews: Our extensive manual inspection of the identified EIRs also revealed that in a tiny fraction (only 30 reviews) the reviewer simply refers to other EIRs to complain about them, indicate his/her awareness and inform other users of such incentivized reviews. However, these reviews are not incentivized themselves. To exclude these reviews, we manually checked random samples of reviews and found that these EIR-aware reviews contain one of the following terms (*who received—with the line "i received—which say they received—their so-called "honest"*).

We then exclude any identified EIR that matched these aware patterns. After extensive manual work in this step, we found only 30 aware reviews by 26 reviewers on 29 products that are excluded from DS2. Interestingly, all these aware reviews were collectively marked as helpful by 194 other users, indicating that many other reviewers felt the same way about the incentivized reviews. This illustrates how the presence of incentivized reviews could impact the trust of customers in the authenticity of Amazon reviews.

4 Basic Characterizations

In this section, we examine a few basic characterizations of EIRs and their associated products and reviewers in order to shed some light on how these elements interact in Amazon.com.

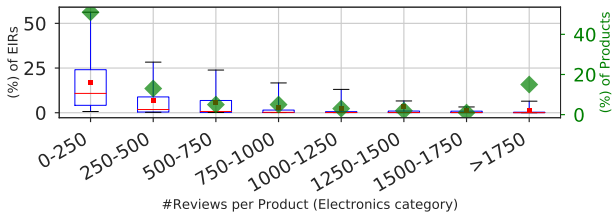


Fig. 1 Distribution of Fraction of EIRs per Product in *Electronics* Category

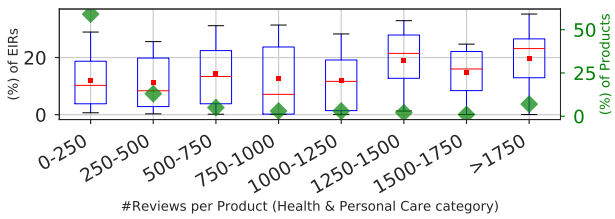


Fig. 2 Distribution of Fraction of EIRs per Product in *Health* Category

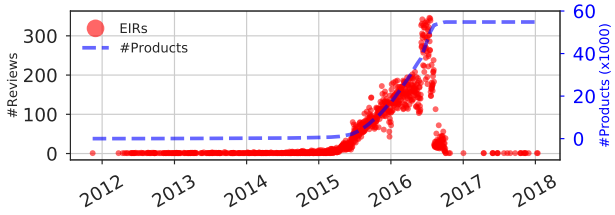


Fig. 3 Evolution of the Daily Number of EIRs and the Total Related Products

4.1 Product Characteristics

One question is *what fraction of reviews for individual products are EIRs?* We use all products in dataset (*DS2*) to examine several characteristics of products that receive at least one EIR.

Fig. 1 and Fig. 2 present the summary distribution of the fraction of product reviews that are EIRs for different groups of products based on the total number of reviews in each category. The red lines (and red dots) show the median (mean) value for each box plot. The green diamonds on these figures show the fraction of all products (per category) that are in each group using the second Y-axis. These figures show that for products in Health and Personal Care category, typically 10-20% of reviews are EIR regardless of the total number of reviews for a product. However, for products in the Electronics category, the fraction of EIRs is generally smaller and rapidly drops as the number of product reviews increases. This suggests that the prevalence of EIRs could vary across different categories of Amazon products.

Another important question is *how the total number of EIR reviews and associated products have changed over time?* Fig. 3 depicts the temporal evolution of the number of observed EIRs per day (with a red dot) as well as the cumulative number of unique products (with the dotted line using the right Y-axis) that received EIRs over time using our DS3 dataset. This figure reveals that while EIRs were present in Amazon at a very low daily rate since 2012, the number of EIRs and associated products have dramatically increased between the middle of 2015 and the middle of 2016. We can clearly observe that Amazon's new policy for banning EIRs (that was announced in October 2016 [1]) have been very effective in rapidly reducing the daily rate of EIRs (and

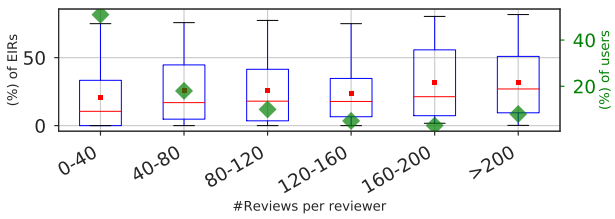


Fig. 4 Distribution of the Fraction of Provided EIRs per Reviewer

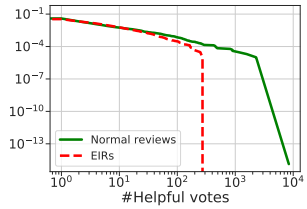


Fig. 5 CCDF of Helpfulness

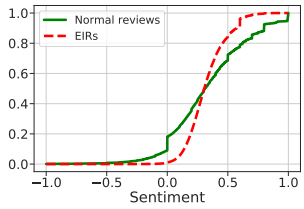


Fig. 6 CDF of Review Sentiment

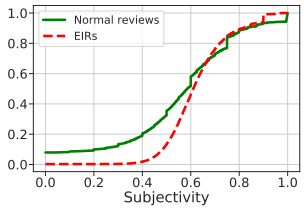


Fig. 7 CDF of Review Subjectivity

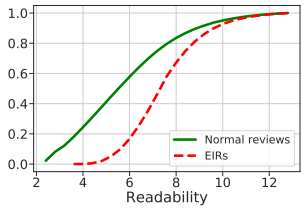


Fig. 8 CDF of Review Readability

the number of affected products) within a couple of months. We note that the effect of this new policy on the implicitly incentivized reviews is unknown.

Another issue is the price range for products that possibly motivate the reviewers to provide EIRs. We observe that 80% (95%) of these products cost less than $25 ($50). In essence, there is typically no significant financial gain in providing a small number of EIRs.

4.2 Reviewer Characteristics

We now turn our attention to reviewers that provided at least one EIR (*i.e.*, reviewers in DS3) to characterize several aspects of these reviewers. We first explore the question of *what fraction of reviews provided by individual reviewers are EIRs?* This illustrates to what extent a reviewer is engaged in writing EIRs.

Fig. 4 presents the summary distribution of the fraction of all reviews of individual users that are EIRs across different groups of users based on their total number of reviews. This figure also presents the number of reviewers in each group (green diamonds) using the second Y-axis. This result illustrates that the fraction of EIRs for most reviewers in DS3 varies between 30-40% of all their reviews. Interestingly, as the reviewers become more active, EIRs make up a more significant fraction of their reviews. To get a better sense of the type (*i.e.*, demography) of users who are likely to provide EIRs, we examined their public profile description and identified the following most common keywords (and their frequencies): "*love*" (1.0) , "*products*" (0.41), "*new*" (0.40), "*Review*" (0.39), "*home*" (0.38), and "*mom*" (0.34).
Our manual inspection of these profiles confirms that around 18% of these reviewers are *moms staying at home that love to review new Amazon products*.

4.3 Review Characteristics

We take a closer look at various features of EIRs in comparison with normal reviews as a reference group.
Helpfulness: An essential aspect of reviews is how helpful they are to other users. Amazon reports the total number of *helpful votes* (up-votes) per review. A slightly larger fraction of normal reviews (12.68%) receive up-votes compare to the EIRs (10.87%). Fig. 5 shows the Complementary Cumulative Distribution Function (CCDF) of the number of up-votes for EIRs and normal reviews. This figure reveals EIRs and normal reviews exhibit the same degree of helpfulness, but the extreme cases for normal reviews are much more helpful.
Review Content: We start by comparing several features of EIR content with normal reviews. First, we observe that 13% of EIRs attach at least one image to their reviews while this ratio is ten times smaller (1.3%) for normal reviews. We perform sentiment analysis on both content and title of reviews using *textblob* library. The sentiment is measured by a value within the range of [-1, 1] where 1 indicates positive, 0 neutral, and -1 a negative sentiment. Fig. 6 presents the distribution of sentiment for the content of EIRs and normal reviews. We observe that 9.5% (9,498) of normal reviews have negative sentiment, 9.1% are neutral (*i.e.*, their sentiment measure is zero) and the rest are positive reviews that are spread across the whole range with some concentration around 0.5, 0.8, and 1. In contrast, the sentiment of nearly all EIRs are positive, but more than 80% of them are between 0 to 0.5. In essence, the sentiment of normal reviews is widespread across the entire range while sentiments for EIRs are mostly positive but more measured. Similarly, less than half of the normal reviews and three-quarter of EIRs have titles with positive sentiments.

Using *TextBlob* library, we also analyzed the *Subjectivity* of reviews, which marks the presence of opinions and evaluations rather than using objective words to provide factual information. Fig. 7 depicts the CDF of the subjectivity across EIRs and normal review datasets. This figure reveals that the subjectivity for 83% of EIRs are between 0.4 and 0.8 while the subjectivity of normal reviews is widely spread across the whole range for normal reviews. We use the *Gunning Fog index* [8], implemented in *textstat* python library[3], to measure the readability test for English writing in each group of reviews. This index estimates the number of years of formal education a person needs to understand the text on the first reading. For example, a Fog index of 12 requires the reading level of a U.S. high school senior. Fig. 8 shows the CDF of the Fog index across EIRs and normal reviews.

[3] https://pypi.org/project/textstat/

This result illustrates that the readability of EIRs requires at least 4 years of education and is 1.5 years higher than normal reviews on average (7.5 vs. 6 years of education). Also, the index exhibits much smaller variations across EIRs. In short, the writing of EIRs is more elaborate.

Self-Similarity of User Reviews: Similarity of the content across submitted reviews by individual users reveals whether a reviewer merely repeat the same set of sentences across different reviews (and thus provides a generic review) or not. To this end, we assess the level of similarity in the text of all pairs of written reviews by a normal reviewer and all pairs of written EIRs by EIR. We use the *Jaccard index* on the uni-grams of reviews as a measure of similarity between the content of a pair of reviews. We consider all reviewers with at least 5 EIRs (1,004 users) from DS2 for this analysis and the same number of randomly selected normal reviewers with the same distribution of reviews as a reference.

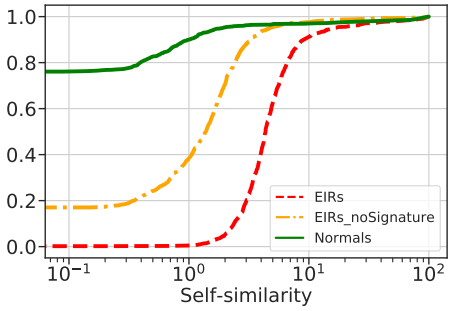


Fig. 9 Pair-wise similarity of the content of submitted reviews by normal and EIR reviewers (with and without the text of disclaimer for their incentive) based on Jaccard index

Fig. 9 depicts the CDF of pairwise similarities between normal reviews and EIRs across normal and EIR reviewers using the log-scale for the x-axis. This figure demonstrates a measurably higher level of self-similarity between EIRs compare to normal reviews. In particular, 75% (97%) of pairs of normal reviews by individual normal reviewers exhibit zero (<5%) similarity. However, 90% (and 39%) of pairs of reviews submitted by EIR reviewers exhibit more than 1% (5%) similarity in their content. Interestingly, we identified 10 EIR and 14 normal reviewers who have submitted roughly between 6 to 10 identical reviews on different products. The presence of the regular expression in EIRs that explicitly indicates the reviewer's incentive could increase the level of similarity between reviews of an EIR reviewer. To ensure that our similarity measure is not significantly affected by the explicitly stated incentive in EIRs, the orange line (labeled EIRs-noSignature) in Fig. 9 also presents the level of similarity between pairs of EIRs per reviewer after removing the identified regular expression from all EIRs. Fig. 9 shows that the level of similarity between reviews of individuals with EIRs is still much larger than reviews of normal reviewers.

Review Submission Timeline: Another interesting question is whether the submission timeline for EIRs vs normal reviews for individual products is different? In particular, over which part of a product lifetime EIRs and normal reviews are submitted. In the absence of any explicit signal, we use the time between the first and last review of a product as an approximation of its lifetime [4] and assign a normalized submission time that we call recency of review for all reviews of products in DS1. For example, recency of 100% indicates that a review was submitted very recently whereas 0% implies a review that is submitted right after a product becomes available. Fig. 10 and 11 show the summary distribution of recency of all reviews (left Y-axis) for different groups of products based on their age in terms of the number of months on the X-axis. Both figures show the fraction of products in each group (right Y-axis). The bottom part of Fig. 10 also presents the prevalence of EIRs in Amazon by showing the number of submitted EIRs in the same window of time. Fig. 11 clearly illustrates that the submission timeline of normal reviews is generally balanced across the life of corresponding products regardless of their age. However, the submission timeline for EIRs exhibits broadly two different patterns based on product age. For products that have become available during the most recent 18-month window when EIRs were prevalent in Amazon, a visibly larger fraction of EIRs was submitted during the first half of product lifetime. However, for older products, the EIRs were submitted during the more recent window since EIRs were not common on Amazon during the first half these products lifetime. In summary, the submission of EIRs in the early part of recent products' lifetime indicates sellers' effort to attract EIRs as they list a new product on Amazon whereas the late submission for older products is simply due to the relatively recent availability of EIRs over products' lifetime.

Length of Reviews: The overall length of a review and its title could be viewed as measures of its level of details. We observe that the typical (*i.e.*, median) length of an EIR (599 characters) is more than three times longer than a normal review (179 characters). Interestingly, the longest normal review (14.8K character) is much longer than the longest EIR (11K character). We observe a similar pattern for the length of reviews based on word count. Furthermore, the title for EIRs are typically 6.6 words long which is two words longer than the title of normal reviews.

Star Rating: A critical aspect of a review is the star rating

[4] Amazon provides the date when a product becomes available for some categories of product. However, we frequently observe cases where a product has multiple versions in the same product page that have become available at different times but share the same pool of reviews. We use the time between the first and last reviews across all versions of a product to deal with this ambiguity in relating specific review to a particular version of a product.

(in the range of 1 to 5 stars) that it assigns to a product. We observe that the assigned rating by EIRs is frequently more positive than normal reviews. More specifically, 95% (75%) of EIRs associated the rating of at least 3 (5) stars while this number drops to 1 (4) for normal reviews.

Statistical Significance: We have shown that several features of EIRs - namely star rating, helpfulness, text and title length, readability, and sentiment - exhibit different distributions compare to normal reviews. This raises the question that whether the reported difference in these distributions are statistically significant. We perform statistical test to answer this question. We observe that none of these features follow a normal distribution as they did not pass the normal test [7]. Therefore, we rely on *Kruskal-Wallis* test [15] to tackle this question. Kruskal-Wallis tests the null hypothesis that the population median of all the groups are equal. If we observe a large p-value (*e.g.*, more than 0.01), then we cannot reject the null hypothesis. We observe that $p < 0.0001$ for the distributions of most of these characteristics and p=0.006 for review helpfulness. These results suggest that the difference between the distribution of these features for EIRs and normal reviews are indeed statistically significant.

Reviewer-Review Mapping Per Product: A majority of reviewers (99.8%) in our EIR dataset (DS2) have written only one EIR for each product. We only found 73 users who have written multiple EIRs for at least one product. These reviews add up to the total of 151 EIRs for 32 unique products. None of the users in our user-centric dataset (DS1) writes multiple EIRs for a single product. Given the one-to-one relationship between the absolute majority of reviewer-review pairs per product, for the rest of our analysis, we assume each reviewer has only a single review per product and vice versa.

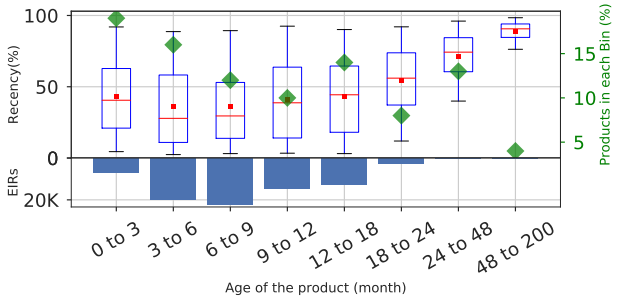


Fig. 10 The submission timeline of EIRs across lifetime of products with different ages

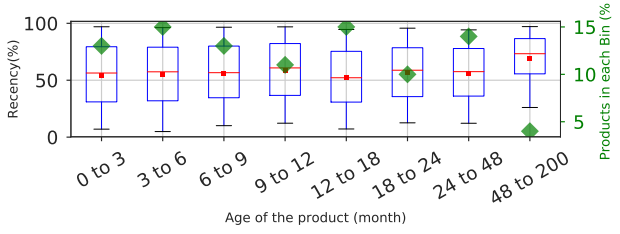


Fig. 11 The submission timeline of normal reviews across lifetime of products with different ages

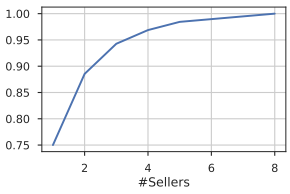


Fig. 12 CDF of the number of sellers that each user submitted EIR for

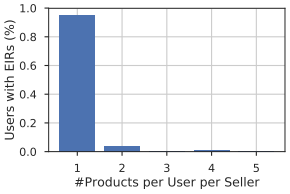


Fig. 13 The distribution of number of products of a seller that each reviewer submits EIR for

Association of Reviewers with Sellers: So far we have primarily focused on the relationship between reviewers and products through reviews. In practice, individual sellers often offer multiple products on Amazon. This raises two questions regarding the associations of EIR reviewers and sellers that we explore here: The first question is *whether an EIR reviewer is typically approached by single or multiple sellers to review their products?* Fig. 12 presents the CDF of the number of unique sellers that each EIR reviewer (in DS3) has submitted at least one EIR for their products. We observe that 75% (95%) of reviewers only submit EIR for at most 1 (3) products of each seller.

The second question is *how many products of a seller a reviewer submit an EIR for?* We examine the distribution of the number of products of a seller that each reviewer in DS3 submits an EIR for. In Fig.13 we observe that the 94.5% (99%) of reviewer-seller relationships are through a single (two) reviewed product(s). *In summary, these results show that EIRs reviewers usually submit an EIR for a single product of one or two sellers.*

Crowdsource Agents: After closer examination of EIRs, we identified hundreds of reviews in DS1 where reviewers explicitly mentioned that they received the products from a specific agency (*e.g.*, *BuzzAgent*, *Influenster*, and *AMZ Review Trader*) in exchange to share their reviews. Our investigations revealed that these websites are associated with crowdsourcing agents that promote a seller's products on different social media platforms. Once a user registers on these websites, she receives certain free products in exchange for her reviews on different social media platforms (*e.g.*, Amazon, Twitter, YouTube). In essence, these websites manage some of the promotional campaigns for selected sellers' products (presumably) for a fee.

Using the name of 12 identified crowdsourcing agencies, we detected 2,124 incentivized reviews from 1,991 reviewers on 237 products among all reviews in DS1. 89.9% of these reviews assigned a strongly positive rating (4 and 5 stars) and 71% of them have a strong positive sentiment. Note that only a small fraction (116 out of 2,124) of these reviews were EIRs. This analysis demonstrates that there are other

indirect ways for sellers to offer incentives to users for submitting positive reviews on Amazon and other more popular social media platforms.

5 Temporal Analysis

All of our previous analysis have focused on the overall characteristics of reviews, reviewers, and products over their entire lifetime. Intuitively, product sellers offer various incentives to attract reviewers and obtain incentivized reviews for their specific product. Obtaining these incentivized reviews over time increases the available information and improves the overall image (*e.g.*, rating) of the product. This, in turn, expands the level of interest among (ordinary) users who may consider to buy the product and provide their own review. Examining the temporal pattern of submitted reviews (by various reviewers) for a product or submitted reviews by a reviewer (for any product) sheds an insightful light in various dynamics among seller products, reviews, and reviewers.

In this section, we tackle two important issues: First, we inspect the *"review profile of sample products"* to study how the temporal pattern of obtained EIRs for a product affect the level of interest among other users. Second, we examine the *"review profile of sample reviewers"* to explore how reviewers get engaged in producing EIRs. To tackle these questions we have inspected temporal patterns for many products and reviewers, and only present a few sample cases to illustrate our key findings better.

In this analysis, we primarily focus on the number of EIRs, non-EIRs (*i.e.*, reviews that are not tagged as EIR by our method) associated with a product (or a reviewer) per day and their (cumulative) average rating. [5] across EIRs or non-EIRs that a product receives or a reviewer assigns.

5.1 Product Reviews

We consider four different products to examine the temporal correlations between the daily number of EIRs and the level of interest among other users, namely the number of non-EIRs and their ratings, for each product.

Note that a product seller can (loosely) control the arrival rate of EIRs by offering incentives (or promotions) with a particular deadline to a specific set of reviewers. We refer to such an event as a *promotional campaign*. The goal of our analysis is to investigate whether and to what extent such a campaign affects the number of non-EIRs and their rating for individual products. Note that a product seller can (loosely) control the arrival rate of EIRs by offering incentives (or promotions) with a certain deadline to a specific set of reviewers. We refer to such an event as a *promotional campaign*. By specifying a deadline for the incentive or promotion, the seller can also force interested users to write their reviews within a specific window of time. We simply assume that any measurable, sudden increase in the number of daily EIRs for a product is triggered, by a promotional campaign that is initiated by its seller. The goal of our analysis is to investigate whether and to what extent such a campaign affects the number of non-EIRs and their rating for individual products. Each plot in Fig. 14 presents the daily number of EIRs (with a red X), the daily number of non-EIRs (with a green diamond), the cumulative average rating for all non-EIR (with dotted green) and EIR (with dotted red lines) for a single product. Each plot also shows the cumulative rating of all reviews with a solid blue line. Three rating lines on each plot are based on the right Y-axis showing the star rating (1 to 5 scale).

Short & Moderately Effective Campaigns: Fig. 14-a shows a product that has been consistently receiving a few daily non-EIR (and not a single EIR) reviews over a roughly two year period. Its average product rating rather consistently drops during 2015. A persistent daily rate of EIR suddenly starts in early 2016 and continues for a few months indicating a likely promotional campaign. The campaign triggers a significant increase in the number of non-EIRs. Interestingly, the average rating of EIR rapidly converges to the average rating of non-EIR (and the overall rating) and not only prevents further dropping but also rather improves the overall rating of this product. This appears to be a short-term (over a few months) and moderately effective promotional campaign by the seller.

Multiple Mild but Ineffective Campaigns : Fig.14 - b presents another product that consistently receives non-EIRs over a one year period. We can also observe ON and OFF periods of EIRs that did not seem to seriously engage other users with this product (*i.e.*, no major increase in the daily rate of non-EIRs). The assigned rating by EIRs is relatively constant, and their gap with the rating of non-EIRs (and overall rating) rapidly grows. Clearly, these multiple mild campaigns are not effective in raising the ratings of the product.

Multiple Intense but Ineffective Campaigns: Fig. 14- c shows a product that has been consistently receiving both EIR and non-EIRs over a year-long period. However, there are two (and possibly three) distinct windows of time (each one is a few weeks long) with pronounced peaks in the number of daily EIRs which suggests two intense campaigns. Interestingly, the first campaign only generates short-term interest among ordinary users (shown as a short-term increase

[5] Amazon appears to rely on some weighted averaging method [5] to calculate the overall rating of a product based on factors such as the recency of a review, its helpfulness and whether it is associated with a verified purchase. Since the details of Amazon's rating method is unknown, we simply rely on a linear moving average of all ratings to determine the overall rating of each product or reviewer over time.

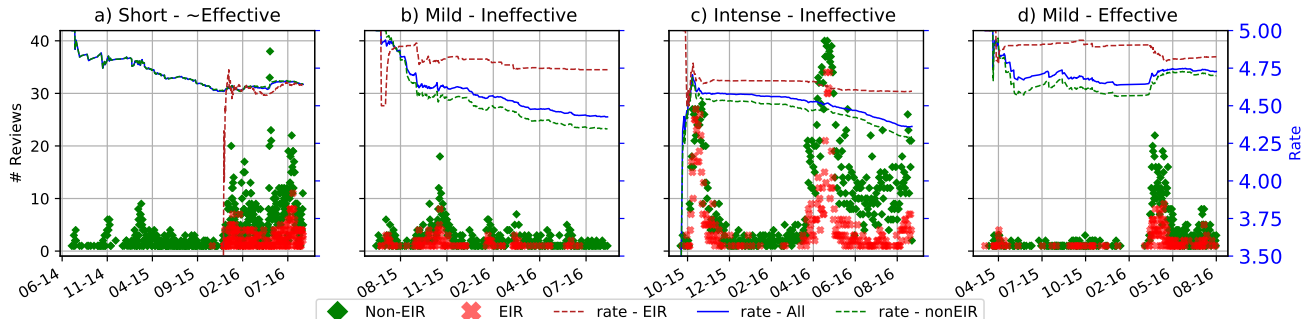


Fig. 14 Temporal Patterns of Reviews for Individual Products

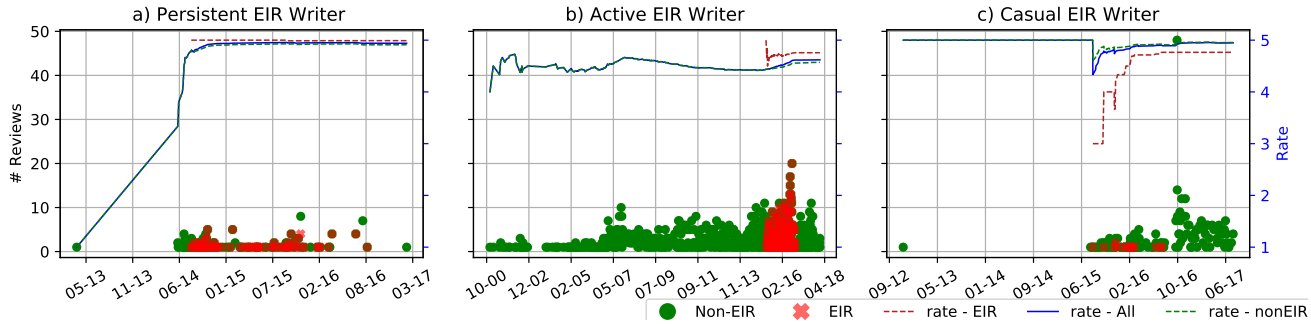


Fig. 15 Temporal Patterns of Reviews for Individual Reviewers

in the daily number of non-EIRs) while the second campaign triggers more non-EIRs. The average rating of EIR is clearly above non-EIRs. However, the average rating of non-EIRs (and even EIRs) continues to drop over time despite the increased level of attention by other users after the second campaign. Therefore, these multiple intense campaigns were not able to improve the overall rating of this product.

Multiple Mild and Effective Campaigns: Fig. 14-d shows a product with a low and persistent daily EIR and non-EIR over a one-year period. We then observe a couple of months with absolutely no reviews that suggest the unavailability of the product. This is followed by a more active campaign of EIRs over a month that continues at a lower rate. This last campaign seems to significantly increase the level of interest among the regular users as well as their rating for this product. In particular, the average rating by non-EIRs was relatively stable and clearly below the rating by EIRs until the last campaign. Interestingly, the last campaign decreases the overall rating by EIRs while enhances the overall rating by non-EIRs. Therefore, we consider this as an effective campaign.

These examples collectively demonstrate that while a seller could loosely control the duration and intensity of its promotional campaign for a product, its impact on the level of engagement by other users could be affected by many other factors (*e.g.*, quality of reviews, strategies of competitors, and quality of the product) and thus widely vary across different products.

5.2 User Reviews

We now focus on the written EIRs and non-EIRs by individual users over time. Similar to the temporal patterns of product reviews, we show the number of daily EIRs (with a red X), non-EIRs (with a green circle). We also show average assigned rating by EIRs (with red dotted line) and non-EIRs (with green dotted line) of the reviewer over time. The three plots in Fig. 15 present the temporal pattern of all reviews (for any product) and their rating for three different reviewers.

Persistent EIR Writer: Fig. 15-a shows a user who provided a single review in 2013 and was inactive for more than a year. Starting in April 2014, she has been submitting a couple of EIRs and/or non-EIRs a day for 20 months, and then her activity significantly dropped. Her average rating for EIRs and non-EIRs are very similar. It appears that this reviewer has become active in Amazon mainly to provide EIRs. But it is rather surprising that she stopped submitting EIR when these types of reviews are very prevalent on Amazon (as we showed in Fig. 3).

Active EIR Writer: Fig. 15-b shows a user who has been actively writing non-EIRs over 16 years since 2001, and her level of activity has gradually increased. Interestingly, she started posting EIRs from 2015 for two years and then stopped. These two years are perfectly aligned with the period in which EIRs have become rapidly popular in Amazon (as we showed in Fig. 3). Furthermore, the overall assigned rating by this reviewer in non-EIRs was relatively stable over time which was slightly lower than her assigned

rating in EIR reviews. This reviewer is a perfect example of a serious Amazon reviewer who takes advantage of offered incentives by sellers for writing EIRs.

Casual EIR Writer: Fig. 15-c shows the temporal pattern of review submission by a user who has been in the system since 2013. However, he became moderately active in the middle of 2015 and provided some EIRs and mostly non-EIRs in the past two years. The number of his EIRs are limited and mostly written over a one year period. It is rather surprising that his rating in EIRs gradually grew over time and was always slightly lower than his ratings for non-EIRs. Far from normal behavior, he has written 49 non-EIRs in one day in 2016 (the green dot above the rating lines). Overall, he appears to be a moderate reviewer who casually writes EIRs.

In summary, our user-level temporal analysis of EIRs and non-EIRs indicates that: *Reviewers exhibit different temporal patterns in producing EIRs. However, users are more active while incentives are offered.*

6 Detecting Other Incentivized Reviews

So far in this paper, we primarily focused on EIRs for our analysis since we can reliably detect and label them as incentivized reviews. However, in practice, there might exist a whole spectrum of explicitly or implicitly incentivized reviews besides EIRs. An intriguing question is *whether all these incentivized reviews (regardless of their implicit and explicit nature) share some common features that can be leveraged to detect them in an automated fashion?* To tackle this question, we consider a number of machine learning and neural network classification methods that are trained using a combination of basic and text features of the reviews.

Pre-processing Reviews: We use 100K random EIRs (from the DS2 dataset) and the same number of normal reviews as our labeled data. First, we remove the sentence that indicates the explicit incentive of a reviewer from each EIR before using the EIRs in this analysis so that these sentences do not serve as a prominent explicit feature. Second, we consider the following pre-processing of text of reviews to examine their exclusive or combined effect on the accuracy of various detection methods: *(i)* converting all characters to lowercase, *(ii)* using the stem of each word in the review (*e.g.*, "wait" is the stem for words "waiting", "waits", "waited") using *Snowball* stemmer of python *NLTK*[6] library. *(iii)* using only alphanumeric characters, and *(iv)* removing all the stop-words using the *NLTK* library in python. (v) converting all the frequent contractions such as *"'ve, 'd, I'm, 'll, n't* to their formal form.

Classification Methods: We examine a number of classification methods including *Multi-Layer Perceptron (MLP)*, *SVM*, *GaussianProcess*, *DecisionTree*, *RandomForest*, *AdaBoost* Classifiers. Each classifier is trained and tested in three scenarios with a different combination of review features as follows: *(i) Basic Features*: Using nine basic features of reviews, length, sentiment, subjectivity, and readability of review text, star-rating, and helpfulness of reviews, as well as length, sentiment, and subjectivity of title, *(ii) Text Features*: Using extracted text features of the review including the word- or character-based {uni, bi, or Tri}-grams (limited to 1500 text features) [7], *(iii) All Features*: Combination of all basic and text-based features. Individual methods are evaluated in 5 and 10-fold cross-validation as well as 70/30 test and training split manner. We only present the result for the 10-fold cross-validation of the MLP method using pre-processed reviews. The results for all other cases are available in our technical report [26].

We found MLP to be considerably better regarding memory usage, computation time, and accuracy on a 50-50% combination of EIR and normal reviews in the training set. Character-based n-grams also led to more effective features compare to word-based n-grams. We use 90% of data for training and testing and 10% of data for hyper-parameter tuning using the *grid-search* in SciKitLearn library. The MLP classifier is trained using default parameters, except for *alpha* (the L2 penalty regularization term) and *hidden_layer_size* that we set to 0.1 and (50,30,10), respectively. Table 2 presents the accuracy, recall, precision, F1-score, Precision-Recall Area Under Curve (P-R AUC), and the Receiver Operating Characteristic (ROC) AUC for MLP Classifier over all runs. These results indicate that even without the explicit acknowledgment sentence in EIRs, a classifier can accurately detect EIRs from normal reviews using basic or text feature. The accuracy further improves if we combine both sets of features.

Model Evaluation: We further evaluate the machine learning model by exploring the logic behind its decision-making process. This exercise demonstrates whether the model is trustworthy and exposes any potential problems in the model that should be addressed. First, we assess how certain our model is in making decisions. Fig. 16 shows the summary

[6] https://www.nltk.org/

[7] We consider character-based n-grams since they are shown to be more robust as they capture spelling differences [12] and are more effective in authorship attribution (writer identification) [14] as they cover a little bit of lexical content, syntactic content, and even style by covering punctuation and white spaces.

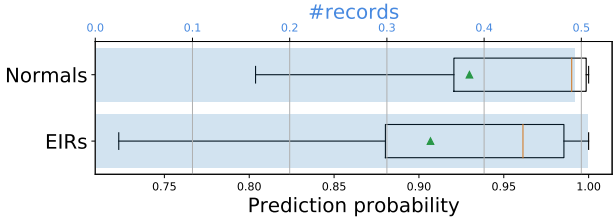


Fig. 16 The classification probability for EIRs and normal reviews along with the number of records in each category

Table 2 The evaluation of MLP classifier in detecting EIRs.

	Accuracy	Recall	Precision	F1-score	Precision-Recall AUC	ROC AUC
Basic	0.85±0.03	0.83±0.01	0.8±0.03	0.83±0.01 (0.84,0.83)	0.88±0.01	0.83±0.01
Text	0.9±0.01	0.89±0.01	0.89±0.01	0.89±0.01 (0.89,0.89)	0.92±0.0	0.89±0.01
Basic+Text	**0.93**±0.02	**0.92**±0.01	**0.92**±0.03	**0.92**±0.01 (0.92,0.92)	**0.94**±0.01	**0.92**±0.01
C-Elect.	0.91±0.03	0.91±0.01	0.91±0.03	0.91±0.01 (0.91,0.91)	0.93±0.01	0.91±0.01
– on Health	0.8	0.87	0.76	0.81 (0.78,0.81)	0.85	0.8
C-Health	0.89±0.03	0.86±0.01	0.83±0.04	0.86±0.01 (0.86,0.86)	0.9±0.01	0.86±0.01
– on Elect.	0.85	0.89	0.83	0.86 (0.85,0.86)	0.89	0.85

distribution of the prediction probability of all test records per class for the model that uses both basic and n-gram features along with the number of records per class (in blue). This result indicates that our model typically exhibits high confidence (92% and 93%) for predicting EIR and normal reviews. However, its confidence has a rather wider variation for EIR records. To explain how our model makes the decisions, we incorporate the LIME [24] framework to assess the feature importance for each of the labels. Fig. 17 and Fig. 18 depicts the summary distribution of feature importance for the top 10 features across all testing samples on predicting EIR and normal reviews, respectively. Features are sorted based on their prevalence (light blue bars). We observe that 8 out of 10 top features are the N-grams of the review content and the other two features are basic overall characteristics of reviews, namely their length and subjectivity. As expected, our model considers reviews with higher subjectivity values as a candidate for EIR and lower ones as normal, although makes this decision in combination with other features.

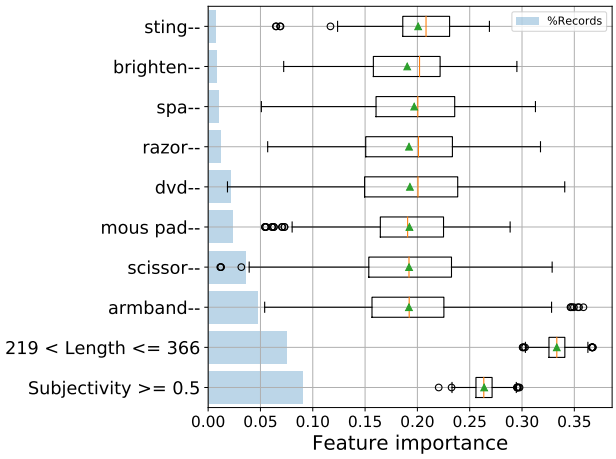


Fig. 17 Importance of top 10 features for EIRs

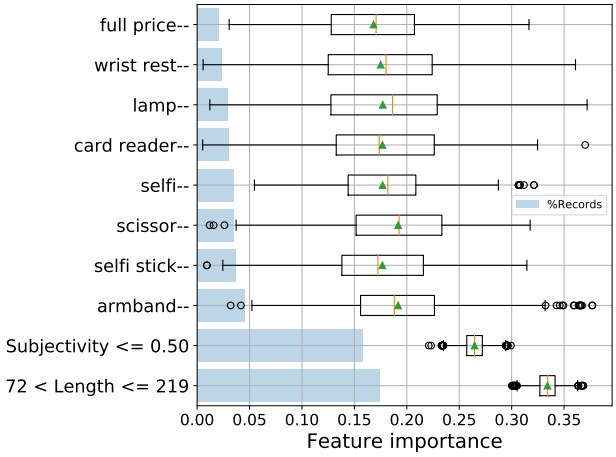


Fig. 18 Importance of top 10 features for Normal Reviews

Category-specific classification: To asses how generally accurate our model can be and whether we need a category specific model, we examine the ability of a classifier for detecting EIRs in other categories. To this end, we divide EIRs and normal reviews into two groups based on the category of their corresponding product (*i.e.*, Electronics and Health). We train two classifiers, called *C-Health* and *C-Elect.* where each one only uses EIRs and normal reviews (with a combination of basic and text features) associated with products in the corresponding category. Finally, we test each classifier on reviews from the same (the 4th and 6th rows of Table 2) as well as on reviews from the other category (the 5th and 7th rows of Table 2) to assess their accuracy in detecting EIR and normal reviews. The last four rows of Table 2 present the accuracy of MLPC for detecting EIRs within each category and between two categories. These results show that the accuracy of detection for EIRs within each category is around 90% and it remains above 80% for cross-category detection of EIRs. Interestingly, the classifier that is trained with Health reviews exhibits a higher accuracy in detecting Electronics reviews.

Next, we investigate the ability of our trained classifier using the basic and text-based features in detecting other incentivized reviews, namely implicitly incentivized reviews (IIRs) and other explicitly incentivized reviews that do not contain the identified regular expressions and thus they were not detected by our method. We randomly select 50,000 re-

views (during 2016) from the DS1 dataset that are neither EIR nor normal reviews. After removing reviews with less than three words in the text, we kept 49,956 reviews. We use the trained classifier to determine whether any of these *unseen* reviews are classified as incentivized or normal reviews. The classifier flags 10,693 (21.4%) of these reviews as incentivized. Our manual inspection of the content of these reviews revealed that they can be broadly divided into two groups as follows:

Other Explicitly Incentivized Reviews: 2,154 (20.1%) of reviews labeled as incentivized contain a variety of different explicit patterns that was so sparse to be captured by our regex, *e.g.*, "I had the opportunity to get it for my review", "received with a promotion rate".

Implicitly Incentivized Reviews (IIRs): We note that the absence of any explicit disclosure of incentives in the remaining reviews does not imply that they are not incentivized. We hypothesize that some of them are implicitly incentivized reviews (IIRs). However, as there is no strong signal to confirm whether these reviews are implicitly incentivized, we consider three different ways to verify our hypothesis as follows:

First, across all the remaining flagged reviews by our classifier, we consider the pairwise relationship between review-product and review-reviewer. We check each of these reviews against the following two conditions: *(i)* whether a review is associated with a product that had received at least one other EIR, or *(ii)* whether a review is provided by a user who has submitted at least one other EIR. We observe that 2,330 (21.8%) reviews are affiliated with both EIR reviewers and EIR products (*i.e.*, meet both conditions) while 3,762 (35.2%) of them are only affiliated with EIR products and 534 reviews are only affiliated with EIR reviewers. Intuitively, meeting both conditions offers stronger evidence that a review could be IIR. Our manual inspection of reviews in these 3 groups confirmed this intuition. While reviews that met both conditions contain an indication of incentive (*e.g.*, *for my honest result*, *promotional price*), reviews related only to products contained moderate hints (*e.g.*, *I have to thank seller*).

Second, we compare the distribution of text and title length, word count, helpfulness, sentiment, subjectivity, readability, and star-rating of 38% of reviews that were labeled as incentivized but neither have explicit pattern nor pairwise relationship with other EIRs. Our analysis show that the distribution of these features for flagged reviews closely follow the corresponding distribution for EIRs, suggesting that these IIRs are flagged correctly. We also use Kruskal-Wallis test to verify the similarity in the distribution of these features for EIRs and flagged reviews. We observe that $p = 0.0$ for all features, except the title length where $p = 0.19$.

Third, since Amazon has warned to remove reviews that violate its guidelines, we examine whether the reviews tagged as IIRs have been removed from Amazon during the past two years. We noticed that 69% of these reviews have been removed from Amazon which indicate the problematic nature of these reviews. We emphasize the presence of the remaining 31% of the reviews does not imply that these are normal reviews.

7 The Current State of EIRs

On October of 2016, Amazon announced the "community guidelines to prohibit incentivized reviews". As we showed in Fig. 3, the daily number of submitted EIRs has significantly dropped after this announcement. In this section, our goal is to examine the status of previously submitted EIRs and related product as well as the ability of sellers to attract incentivized reviews.

We check the availability of more than half of randomly selected EIRs in DS2 in December of 2018, more than two years after we originally collected them. Our analysis showed that an absolute majority of these EIRs (98.5%) were removed from Amazon. Furthermore, when we tried to submit a review for a product that recently had a promotional campaign, we received the message shown in Fig 19 indicating the limitation to submit any reviews for this product due to unusual reviewing activity. This suggests that Amazon has actively limited the submission of new reviews for these problematic products.

We also inspected the content of the tiny fraction of remaining EIRs in the system and noticed that the content of roughly 20% of them was modified. Interestingly, half of these modified EIRs were shortened by removing the disclaimer of explicit incentive by the reviewer. This could be the reason that they were not removed. The other half of modified EIRs were short reviews that later added some personal experience (*e.g.*, what is in the box, how to use the product, pros, and cons).

We performed the following experiment to verify whether and how Amazon sellers might attract incentivized reviews from users. To this end, we made a purchase on Amazon and received the product that included seller's instructions to contact them for receiving other products for a free or discounted price as a VIP customer. We emailed the customer service of this (and a few other) seller(s) indicating

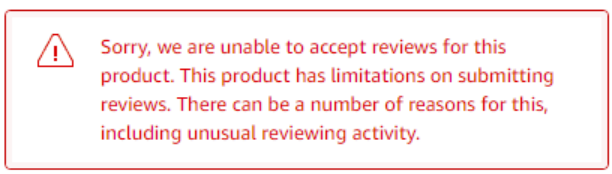
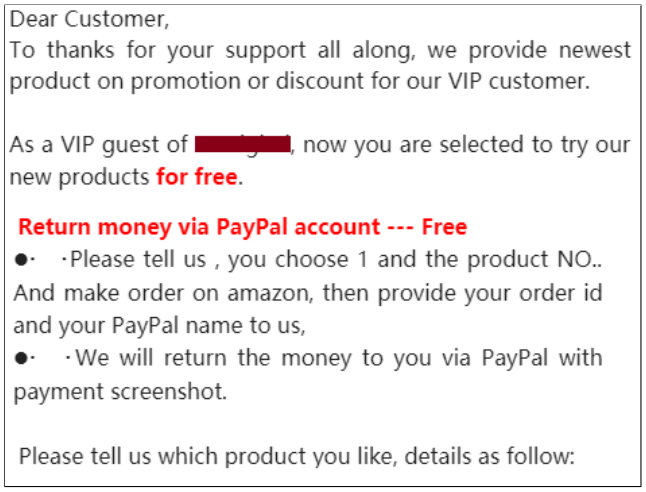


Fig. 19 Amazon limits review submission for products with suspicious review activity


Dear Customer,
To thanks for your support all along, we provide newest product on promotion or discount for our VIP customer.

As a VIP guest of [redacted], now you are selected to try our new products **for free**.

Return money via PayPal account --- Free

●· ·Please tell us , you choose 1 and the product NO.. And make order on amazon, then provide your order id and your PayPal name to us,

●· ·We will return the money to you via PayPal with payment screenshot.

Please tell us which product you like, details as follow:

Fig. 20 Sample of a seller response to users who are interested in becoming a VIP customers that provides instructions for submitting incentivized reviews.

our interest to be a VIP customer and received a variant of the response showed in Fig.20. This response basically provides the instructions for users to buy a product on Amazon, submit their (incentivized) review, and then be reimbursed for their purchase through their Paypal account, *i.e.*, receiving a product for free. Clearly, such an incentivized review does not need to contain any disclaimer and would not be detectable by Amazon. This is a clear indication that incentivized reviews still exist on Amazon but they are not explicit since exchanged information and money between sellers and reviewers are not visible to Amazon.

8 Related Work

Detection and analyzing of spam reviews started in 2008 by labeling the (near) duplicate reviews as spam and using supervised learning techniques to detect spam reviews [10]. Since then, different aspects of online reviews have been investigated such as behavioral abnormalities of reviewers [17] and review quality and helpfulness [19,13,18]. Studies on spam detection have deployed a diverse set of techniques. Early studies relied on unexpected class association rules [11] and standard word and part of speech n-gram features with supervised learning [20] that are later improved by using more diverse feature sets [16]. *FraudEagle* [2] was proposed as a scalable and unsupervised framework that formulates opinion fraud as a network classification problem on a signed network of software product reviews of an app store. These studies also relied on different strategies, such as Amazon Mechanical Turk [20] or manual labeling [16] to create a labeled dataset for their analysis.

The effect of incentives on reviewers and quality of reviews are studied by Qiao et al. [22]. They showed that external incentives might implicitly shift an individuals decision-making context from a pro-social environment to an incentive-based environment. Wang et al. [27] modeled the impact of bonus rewards, sponsorship disclosure, and choice freedom on the quality of paid reviews. In a qualitative study, Petrescu et al. [21] examined the motivations behind incentivized reviews as well as the relationship between incentivized reviews and the satisfaction ratings assigned by consumers to a product. They showed that the level of user engagement depends on a cost-benefit analysis. Burtch et al. [6] focused on social norms instead of financial incentives. By informing individuals about the volume of reviews authored by peers, they test the impact of financial, social norms, and a combination of both incentives in motivating reviewers. The study by Xie [28] unveiled the underground market for app promotion and statistically analyzed the promotion incentives, characteristics of promoted apps and suspicious reviewers in multiple app review services.

To the best of our knowledge, none of the prior studies have systematically examined the prevalence of EIRs, their basic characteristics, and their influence on the level of interest among other users to a product based on large-scale quantitative measurements in a major e-commerce platform.

9 Conclusion

In this paper, we presented a detailed characterization of Explicitly Incentivized Reviews (EIRs) in two popular categories of Amazon products. We presented a technique to detect EIRs, collected a few datasets from Amazon and identified a large number of EIRs in Amazon along with their associated product and reviewer information. Using this information, we compared and contrasted various features of EIRs with reasonably normal reviews. We showed that EIRs exhibit different features compared to normal reviews and discussed the implications of these differences.

Then, we zoomed into the temporal pattern of submitted EIR reviews for a few specific products and submitted reviews by a few specific reviewers. These temporal dynamics demonstrated whether/how promotional campaigns by a seller could affect the level of interest by other users and how reviewers could get engaged in providing EIRs. Finally, we illustrated that machine learning techniques can identify EIRs from normal reviews with a high level of accuracy.

Moreover, such techniques can accurately identify other explicitly and implicitly incentivized reviews. We leverage affiliation of reviews with reviewers and products to infer their incentivized nature.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grants No. CNS-1564348 and

CHS-1551817. We gratefully acknowledge the support of Intel Corporation for giving access to the Intel AI DevCloud platform used for this work.

References

1. aboutAmazon.com. Update customer review, https://goo.gl/fiVa8j, 2016.
2. L. Akoglu, R. Chandy, and C. Faloutsos. Opinion fraud detection in online reviews by network effects. *Proceedings of the ICWSM*, 2013.
3. Amazon.com. Community guidelines, https://www.amazon.com/gp/help/customer/display.html? nodeid=14279631, 2018.
4. Amazon.com. About amazon verified purchase reviews, https://goo.gl/aNcPCR, 2018.
5. T. Bishop. https://www.geekwire.com/2015/amazon-changes-its-influential-formula-for-calculating-product-ratings/, 2015.
6. G. Burtch, Y. Hong, R. Bapna, and V. Griskevicius. Stimulating online reviews by combining financial incentives and social norms. *Management Science*, 2017.
7. R. B. d'Agostino. An omnibus test of normality for moderate and large size samples. *Biometrika*, 58(2):341–348, 1971.
8. R. Gunning. The technique of clear writing. *McGraw-Hill, NY*, 1952.
9. S. Jamshidi, R. Rejaie, and J. Li. Trojan horses in amazon's castle: Understanding the incentivized online reviews. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 335–342, 2018.
10. N. Jindal and B. Liu. Opinion spam and analysis. In *ACM International Conference on Web Search and Data Mining*, 2008.
11. N. Jindal, B. Liu, and E.-P. Lim. Finding unusual review patterns using unexpected rules. In *Proceedings of the ACM international Conference on Information and Knowledge Management*, 2010.
12. I. Kanaris, K. Kanaris, I. Houvardas, and E. Stamatatos. Words versus character n-grams for anti-spam filtering. *International Journal on Artificial Intelligence Tools*, 16(06):1047–1067, 2007.
13. S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti. Automatically assessing review helpfulness. In *Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing*, 2006.
14. M. Koppel, J. Schler, and S. Argamon. Authorship attribution in the wild. *Language Resources and Evaluation*, 45(1):83–94, 2011.
15. W. H. Kruskal and W. A. Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.
16. F. Li, M. Huang, Y. Yang, and X. Zhu. Learning to identify review spam. In *Proceedings of IJCAI*, 2011.
17. P. Lim, V. Nguyen, N. Jindal, B. Liu, and H. Lauw. Detecting product review spammers using rating behaviors. In *Proceedings of ACM International Conference on Information and Knowledge Management*, 2010.
18. J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou. Low-quality product review detection in opinion summarization. In *Proceedings of the Joint Conference on EMNLP-CoNLL*, 2007.
19. S. Mudambi. What makes a helpful online review? a study of customer reviews on amazon.com. *MIS Quarterly*, 34:185–200, 2010.
20. M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the ACL Human Language Technologies*, 2011.
21. M. Petrescu, K. OLeary, D. Goldring, and S. B. Mrad. Incentivized reviews: Promising the moon for a few stars. *Journal of Retailing and Consumer Services*, 2017.
22. D. Qiao, S.-Y. Lee, A. Whinston, and Q. Wei. Incentive provision and pro-social behaviors. In *Proceedings of the Hawaii International Conference on System Sciences*, 2017.
23. ReviewMeta.com. Analysis of 7 million amazon reviews, https://goo.gl/CPzHpB, 2016.
24. M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM, 2016.
25. K. Shyong, D. Frankowski, J. Riedl, et al. Do you trust your recommendations? an exploration of security and privacy issues in recommender systems. *Emerging Trends in ICS*, 2006.
26. J. Soheil, R. Reza, and L. Jun. Characterizing the incentivized online reviews. *Technical Report , University of Oregon*, 2016-18, URL: https://www.cs.uoregon.edu/Reports/TR-2018-001.pdf.
27. J. Wang, A. Ghose, and P. Ipeirotis. Bonus, disclosure, and choice: What motivates the creation of high-quality paid reviews? In *Proceedings of the International Conference on Information Systems*, 2012.
28. Z. Xie and S. Zhu. Appwatcher: Unveiling the underground market of trading mobile app reviews. In *Proceedings of the ACM Conference on Security & Privacy in Wireless and Mobile Networks*, 2015.