Trojan Horses in Amazon's Castle: Understanding the Incentivized Online Reviews

Soheil Jamshidi, Reza Rejaie, Jun Li
Department of Computer and Information Science
University of Oregon
{Jamshidi, Reza, Lijun}@cs.uoregon.edu

***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. Finally, we 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. Overall, this analysis sheds an insightful light on the impact of EIRs on Amazon products and users.

I. Introduction

As the popularity of online shopping has rapidly grown during the past decade, the shoppers have increasingly relied on the 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 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 [7], [11], [8], [14], [10], [2] and also biased and paid reviews [18], [20], [21], [15], [5] 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. We refer to these reviews as *incentivized reviews*. 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 [17], 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 under cover.

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/subcategories and collect all the information for top-20 best-seller products in all subcategories of two major categories. The first contribution of this paper is a method to identify explicitly incentivized reviews (EIRs) on Amazon. We identify 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.

The second contribution of this paper is the characterization of 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)*

IEEE/ACM ASONAM 2018, August 28-31, 2018, Barcelona, Spain
978-1-5386-6051-5/18/$31.00 © 2018 IEEE

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. Finally, given an apparent gap between 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 incentivized reviews.

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

II. 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 sub-categories. 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.

TABLE I
Basic Features of Our Datasets

	Products (DS1)	EIRs (DS2)	Normal Reviews	Reviewers (DS3)
Reviews	3,797,575	100,086	100,086	217,000
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 III, 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 and only consider their reviews for our analysis.

The DS1, DS2, and Normal reviews datasets were collected in December 2016, and the Reviewers dataset (DS3) was collected in January 2018.

III. 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

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

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 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:

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

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 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.

IV. 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.

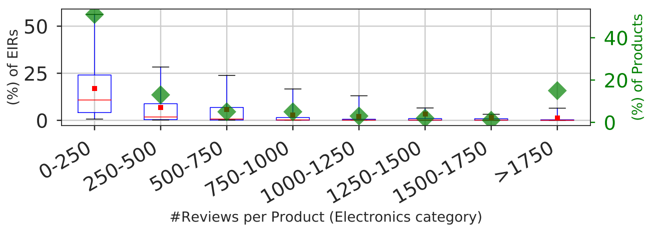

Fig. 1. Distribution of Fraction of EIRs per Product in *Electronics* Category

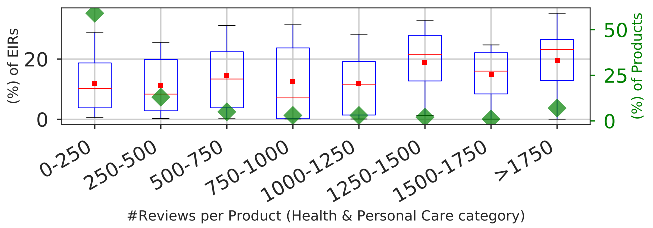

Fig. 2. Distribution of Fraction of EIRs per Product in *Health* Category

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 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

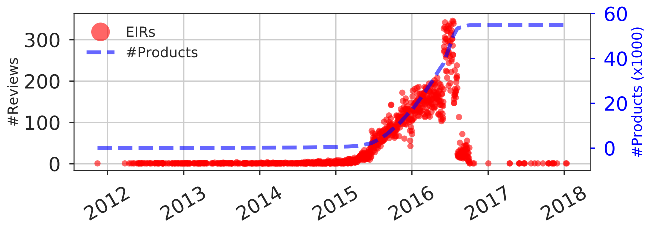

Fig. 3. Evolution of the Daily Number of EIRs and the Total Related Products

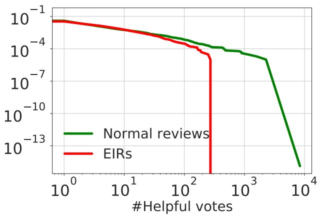

Fig. 4. CCDF of Helpfulness

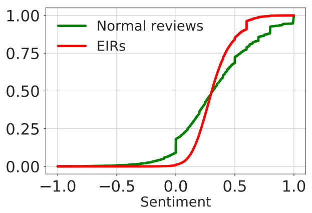

Fig. 5. CDF of Review Sentiment

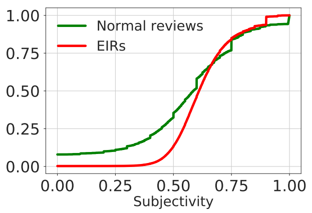

Fig. 6. CDF of Review Subjectivity

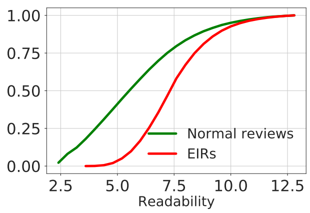

Fig. 7. CDF of Review Readability

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 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.

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. 8 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 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*.

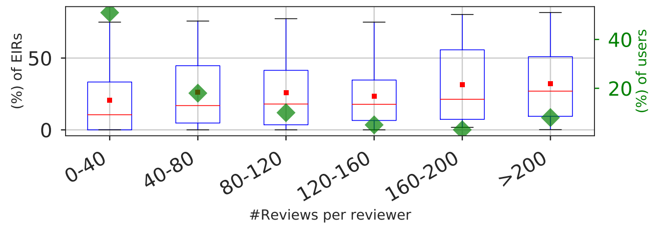

Fig. 8. Distribution of the Fraction of Provided EIRs per Reviewer

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. 4 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. 5 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. 6 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* [6] to measure the readability test for English writing

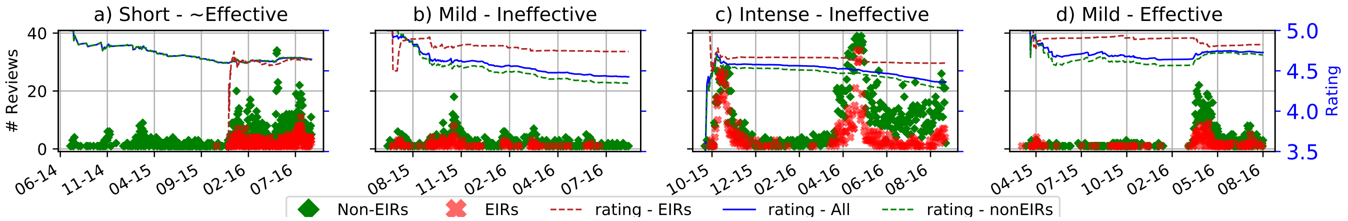

Fig. 9. Temporal Patterns of Reviews for Individual Products

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. 7 shows the CDF of the Fog index across EIRs and normal reviews. 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.

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 (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.

Reviewer-Review Mapping Per Product: A majority (99.8%) of reviewers 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.

V. 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 products. Obtaining these incentivized reviews over time increases the available information and improves the overall image (*e.g.*, rating) of a 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 product sellers, 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 affects 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 that better illustrate our key findings.

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. Each plot in Fig. 9 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 a dotted green line) and EIR (with a dotted red line) 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. 9-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 slightly 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 EIRs rapidly converges to the average rating of non-EIRs (and the overall rating) and not only prevents further dropping but also slightly 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. 9-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 rating of the product.

Multiple Intense but Ineffective Campaigns: Fig. 9-c shows a product that has been consistently receiving both EIR and non-EIRs over a year-long period. However, there are two 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 in the daily number of non-EIRs) while the second campaign triggers a longer term increase in 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 engagement by regular 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. 9-d shows a product with a low and persistent daily EIR and non-EIR over a one-year period. We 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 it enhances the overall rating by non-EIRs. Therefore, we consider this 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 regular users and their rating could be affected by many other factors (*e.g.*, quality of reviews and product, strategies of competitors, and product rank on different search queries) and thus widely varies across different products.

User Reviews

We now focus on the submitted 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), and non-EIRs (a green circle). We also show average assigned rating of the submitted EIRs (red dotted line) and non-EIRs (green dotted line) of the reviewer over time. The two plots in Fig. 10 present the temporal pattern of all reviews (for any product) and their rating for two different reviewers.

Active EIR Writer: Fig. 10-a shows a user who has been actively writing non-EIRs over 17 years since 2001, and her level of activity has gradually increased. Interestingly, she started posting EIRs from 2015, continued 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 EIRs. This reviewer is a perfect example of a serious Amazon reviewer who takes advantage of offered incentives by sellers for writing EIRs.

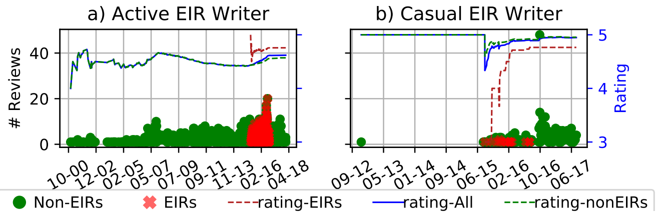

Fig. 10. Temporal Patterns of Reviews for Individual Reviewers

Casual EIR Writer: Fig. 10-b 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 in submitting EIRs when incentives are offered.*

VI. 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 DS3 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 lower-case, *(ii)* using the stem of each word in the review (*e.g.*, “wait” is the stem for words “waiting”, “waits”, “waited”). *(iii)* using only alphabet characters, and *(iv)* removing all the stop-words.

Classification Methods: We examine a number of classification methods including *Multi-Layer Perceptron (MLP)*, *SVM*, *GaussianProcess*, *DecisionTree*, *RandomForest*, *AdaBoost* Classifiers, *Bi-grams* and *Tri-grams* (with and without *tf-idf*), and *character-based bi- and tri-grams*. 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 character-based Tri-grams (limited to 2**10 text features) of the reviews, *(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 MLP method using pre-processed reviews after removing all stop words and replacing all remaining words with their stem part as this combination exhibits the highest level of accuracy. The results for all other cases are available in our technical report [19].

We found MLPC 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. 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 II presents the average 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.

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 one category. Finally, we test each classifier on reviews from the other category to assess their accuracy in detecting EIR and normal reviews. The last two rows of Table II present the accuracy of MLPC for this cross-category detection of EIRs. These results show that the accuracy of cross-category detection of EIRs (for these two categories) is still sufficiently high ($\geq$80%). Interestingly, the classifier that is trained with Health reviews exhibits a higher accuracy in detecting Electronics reviews.

TABLE II
THE EVALUATION OF MLP CLASSIFIER IN DETECTING EIRS.

	Acc.	Rec.	Prec.	F1-score	P-R AUC	AUC
Basic	0.84	0.81	0.78	0.81	0.86	0.81
Text	0.88	0.89	0.89	0.89	0.91	0.89
Basic+Text	**0.92**	**0.89**	0.86	**0.89**	**0.93**	**0.89**
C-Elect.	0.8	0.8	0.79	0.8	0.85	0.8
C-Health	0.87	0.86	0.84	0.86	0.9	0.86

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 100,000 reviews (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 98,594 reviews. We use the trained classifier to determine whether any of these *unseen* reviews are classified as incentivized or normal reviews. The classifier flags 20,892 (21.19%) 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: 3,799 (18%) of reviews labeled as incentivized contain a variety of different explicit patterns that was hard 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). To verify this hypothesis across all the remaining flagged reviews, we rely on the pairwise relationship between review-product and review-reviewer and check any 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 296 (1.4%) reviews are affiliated with both EIR reviewers and EIR products (*i.e.*, meet both conditions) while 8544 (41%) of them are only affiliated with EIR products and 63 reviews are only affiliated with EIR reviewers. Intuitively, meeting both conditions offers a 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 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*).

VII. 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 [7]. Since then, different aspects of online reviews have been investigated such as behavioral abnormalities of reviewers [11] and review quality and helpfulness [13], [9], [12]. Studies on spam detection have deployed a diverse set of techniques. Early studies relied on unexpected class association rules [8] and standard word and part of speech n-gram features with supervised learning [14] that are later improved by using more diverse feature sets [10]. *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 [14] or manual labeling [10] to create a labeled dataset for their analysis.

The effect of incentives on reviewers and quality of reviews are studied by Qiao et al. [16]. 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. [20] modeled the impact of bonus rewards, sponsorship disclosure, and choice freedom on the quality of paid reviews. In a qualitative study, Petrescu et al. [15] 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. [5] 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 [21] 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.

VIII. 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. We then 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.

Some of our future plans are as follows: We plan to iteratively improve the performance of classifiers by incorporating other explicit patterns. Furthermore, we deploy probabilistic techniques to infer the likelihood that a review is incentivized based on its affiliation with other products and reviewers. Finally, we explore whether the incentivized reviews have disappeared entirely from Amazon or become more implicit.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grants No. CNS-1564348 and CHS-1551817.

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. *Proc. 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] G. Burtch, Y. Hong, R. Bapna, and V. Griskevicius. Stimulating online reviews by combining financial incentives and social norms. *Management Science*, 2017.

[6] R. Gunning. The technique of clear writing. *McGraw-Hill, NY*, 1952.

[7] N. Jindal and B. Liu. Opinion spam and analysis. In *ACM International Conference on Web Search and Data Mining*, 2008.

[8] N. Jindal, B. Liu, and E.-P. Lim. Finding unusual review patterns using unexpected rules. In *Proc. of the ACM international Conference on Information and Knowledge Management*, 2010.

[9] S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti. Automatically assessing review helpfulness. In *Proc. of the ACL Conference on Empirical Methods in Natural Language Processing*, 2006.

[10] F. Li, M. Huang, Y. Yang, and X. Zhu. Learning to identify review spam. In *Proc. of IJCAI*, 2011.

[11] P. Lim, V. Nguyen, N. Jindal, B. Liu, and H. Lauw. Detecting product review spammers using rating behaviors. In *Proc. of ACM International Conference on Info. and Knowledge Management*, 2010.

[12] J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou. Low-quality product review detection in opinion summarization. In *Proc. of the Joint Conference on EMNLP-CoNLL*, 2007.

[13] S. Mudambi. What makes a helpful online review? a study of customer reviews on amazon.com. *MIS Quarterly*, 34:185–200, 2010.

[14] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proc. of the ACL Human Language Technologies*, 2011.

[15] 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.

[16] D. Qiao, S.-Y. Lee, A. Whinston, and Q. Wei. Incentive provision and pro-social behaviors. In *Proc. of the Hawaii International Conference on System Sciences*, 2017.

[17] ReviewMeta.com. Analysis of 7 million amazon reviews, https://goo.gl/CPzHpB, 2016.

[18] 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.

[19] J. Soheil, R. Reza, and L. Jun. Characterizing the incentivized online reviews. *Technical Report , University of Oregon*, 2016-18.

[20] J. Wang, A. Ghose, and P. Ipeirotis. Bonus, disclosure, and choice: What motivates the creation of high-quality paid reviews? In *Proc. of the International Conference on Information Systems*, 2012.

[21] Z. Xie and S. Zhu. Appwatcher: Unveiling the underground market of trading mobile app reviews. In *Proc. of the ACM Conference on Security & Privacy in Wireless and Mobile Networks*, 2015.