Review Trade: Everything Is Free in Incentivized Review Groups

Yubao Zhang[1], Shuai Hao[2(✉)], and Haining Wang[3]

[1] University of Delaware, Newark, DE, USA
ybzhang@udel.edu
[2] Old Dominion University, Norfolk, VA, USA
shao@odu.edu
[3] Virginia Tech, Arlington, VA, USA
hnw@vt.edu

Abstract. Online reviews play a crucial role in the ecosystem of e-commerce business. To manipulate consumers' opinions, some sellers of e-commerce platforms outsource opinion spamming with incentives (*e.g.*, free products) in exchange for *incentivized reviews*. As incentives, by nature, are likely to drive more biased reviews or even fake reviews. Despite e-commerce platforms such as Amazon have taken initiatives to squash the incentivized review practice, sellers turn to various social networking platforms (*e.g.*, Facebook) to outsource the incentivized reviews. The aggregation of sellers who request incentivized reviews and reviewers who seek incentives forms *incentivized review groups*. In this paper, we focus on the incentivized review groups in e-commerce platforms. We perform data collections from various social networking platforms, including Facebook, WeChat, and Douban. A measurement study of incentivized review groups is conducted with regards to group members, group activities, and products. To identify the incentivized review groups, we propose a new detection approach based on co-review graphs. Specifically, we employ the community detection method to find suspicious communities from co-review graphs. Also, we build a "gold standard" dataset from the data we collected, which contains the information of reviewers who belong to incentivized review groups. We utilize the "gold standard" dataset to evaluate the effectiveness of our detection approach.

Keywords: Incentivized review groups · Co-review graph · Community detection

1 Introduction

Online reviews on commercial products and services extensively impact consumers' decision making. As reported, 90% of consumers read online reviews before purchasing a product or service, and 88% of consumers tend to trust online reviews as much as personal recommendations [3]. About 80% of consumers reverse the decisions of product purchase after reading negative reviews, and 87% of consumers affirm a purchase decision based on positive reviews [9].

N. Park et al. (Eds.): SecureComm 2020, LNICST 335, pp. 339–359, 2020.
https://doi.org/10.1007/978-3-030-63086-7_19

Therefore, today's merchants are strongly motivated to fabricate online reviews in order to manipulate custom opinions. One of the most popular ways for fabricating positive reviews is called *incentivized reviews*, *i.e.*, merchants bribe reviewers by providing free products or even offer compensation for favorable reviews (*e.g.*, five-star reviews on Amazon). With incentivized reviews, merchants could gain a competitive advantage over rival merchants, as customers prefer online products with a larger number of favorable reviews.

To further affect people's thoughts and decisions, incentivized reviews nowadays are collected from a group of reviewers (*i.e.*, the *incentivized review groups*) so as to perform opinion spamming. In particular, incentivized review groups are online venues for trading reviews, where merchants can post the products that seek to favorable reviews and reviewers can write such reviews to earn free products or even make extra compensation. Some of the merchants designate well-written reviews to reviewers such that they can guarantee the quality of incentivized reviews. As such, there emerges a shady business that acts as a go-between of merchants and consumers, such as review outsourcing websites.

Apparently, the underground industry of fabricating fake reviews mentioned above violates the rules of most e-commerce platforms. As Amazon's consumer review policy [2] states, the violations include "a seller posts a review of their own product or their competitor's product" and "a seller offers a third party a financial reward, discount, free products, or other compensation in exchange for a review", *etc.* Despite the strict prohibition of Amazon (*i.e.*, banning/closing the accounts of both merchants and consumers), incentivized review groups are still thriving across different platforms, especially the social media such as Facebook and WeChat. This shady industry produces a spate of fake reviews, which mislead the customers, damage the trust of reviews, and even endanger the healthiness of the e-commerce ecosystem.

In this paper, we investigate the incentivized review groups on Amazon, the most popular e-commerce platform. To understand the breadth of the problem, we collect incentivized review groups across several different platforms, including Facebook, WeChat, and Douban. We find that different platforms play different roles in the ecosystem of incentivized review groups. Specifically, incentivized review groups on Facebook act like the blackboards, where a set of merchants (*i.e.*, sellers) post their products directly in these Facebook groups. Meanwhile, incentivized review groups on Douban are of the service for merchants and brokers, which educate them how to effectively obtain incentivized reviews. The incentivized review groups on WeChat are most private and generally are owned by a single person, who recruits reviewers to join the group and posts review requests for a set of products.

To understand the incentivized review groups, we conduct a measurement study to collect and characterize real review groups. We investigate the number and the increment rate of review members, as well as the number of merchants in collected incentivized review groups. In terms of incentivized review requests, we inspect the incentivized review requests in different groups as well as from

Fig. 1. Amazon incentivized review group

individual merchants. We also examine the categories, questions & answers, and the relationship between merchant and manufacturers of products.

Based on the measurement study, we then propose a graph-based method to detect the incentivized review groups on Amazon by leveraging the co-review behavior among reviewers. *Co-review* reflects the actions that two reviewers post reviews on the same product. By constructing the co-review graphs of reviewers, we then employ the community detection method to identify suspicious communities. Specifically, we consider both the frequency of co-reviews and other important features of the co-review behavior, such as co-reviews in a burst, which could significantly imply the existence of incentivized reviews. Furthermore, we note that the detection of incentivized review groups can be further integrated into the existing spam review detection framework to improve its coverage and effectiveness.

To evaluate our detection method, we construct a "gold standard" dataset by collecting real review requests from popular incentivized review groups, which enables us to validate the effectiveness of our method and shed light on further fake review studies (our dataset has been made publicly available at https://github.com/zhangyubao5/incentivized_review_group). Then, we examine an extensive Amazon review dataset [10,17] ranging from 1996 to 2018 and find that incentivized review groups started to pose serious threats on the ecosystem of online reviews after 2014.

2 Background

Obtaining positive reviews is one major factor of being successful online sellers. When competing with similar products at a similar price, the product with a higher rate or better reviews is more likely to win out.

2.1 Incentivized Reviews

To obtain positive reviews in the short term, sellers provide free products or offer compensation, *i.e.*, the "*incentivized reviews*". With the incentive for reviewers, it is guaranteed that sellers can obtain positive reviews (such as five-star in Amazon) and enhance the rating or recommendation of the products expeditiously. However, incentivized reviews typically violate the policy of online platforms since they are published in exchange for free products or compensation.

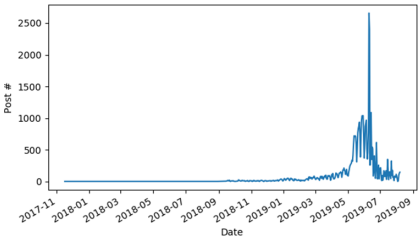


Fig. 2. Facebook review groups.

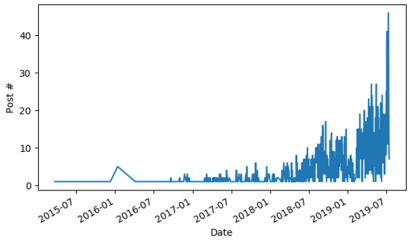


Fig. 3. Douban review groups.

For example, Amazon announced the policy to ban the incentivized reviews in 2016 [1].

2.2 Verified Purchase

Around the same time when Amazon started the crackdown on incentivized reviews, Amazon introduced the "verified purchase" tag. A "verified purchase" tag is labelled with the customer review if Amazon can verify that the review was published by the account who made the purchase. Although the "verified purchase" tag can highlight some authentic reviews and hinder the spam reviews to a certain degree, crooked sellers can bypass the hurdle or even exploit the "verified purchase" through review groups.

2.3 Incentivized Review Group

Incentivized review groups, or incentivized review clubs, are communities created to connect the consumers who want free products or compensation and the sellers who want positive product reviews. Figure 1 illustrates how the incentivized review groups work. First, a seller posts the products that need reviews and reviewers register for particular products of their interest. After the registration is confirmed by the sellers, buyers purchase the products in Amazon and write favorable reviews after the completion of orders. Then, they would show the proof of favorable reviews to the seller and obtain reimbursement or compensation. The registration enables the seller to follow up and ensure that the buyers have posted the reviews and the reviews are favorable.

Since buyers make payments on Amazon at full price, they are eligible for posting "verified purchase" reviews. Once the reviews have been confirmed, sellers send the cost of their purchases back, sometimes with extra compensation. Despite Amazon's strict policy against incentivized reviews, a number of incentivized review groups are still operating on websites or social media platforms. There are a great number of incentivized review groups on Facebook, which are set up specifically for Amazon sellers. Incentivized review groups usually set their groups as private or require sign-up to view the posts. Some of them claim the rules of incentivized review groups, including no scam, no hate speech,

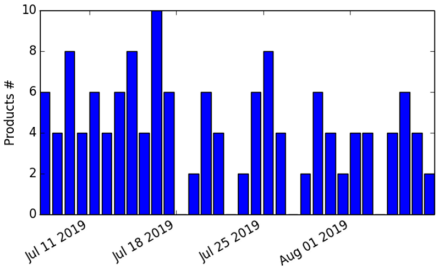


Fig. 4. A review group in WeChat.

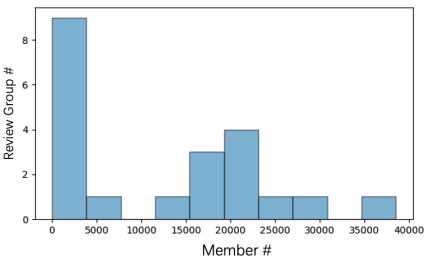


Fig. 5. Number of group members.

and encouraging users to report invalid posts. Sellers also operate the incentivized review groups on other websites (such as Douban and Reddit) or instant messaging applications (such as WeChat).

3 Data Collection

In this section, we describe the data collection of incentivized review groups and summarize the datasets. We collect incentivized review groups from various social networking platforms, including Facebook, WeChat, and Douban. WeChat is the most popular instant messaging application in China, which allows users to create groups to broadcast events to group members. Douban is one of the most influential social networking service website in China, which allows users to create interest groups to share information.

3.1 Dataset

Facebook: There are many Facebook groups that are abused by incentivized review groups. Some of them are private and only allow group members to view the posts. To obtain good coverage and representativeness, we select 20 most active public incentivized review groups on Facebook, where we observe multiple tasks each day. In addition, we sign up two private groups by sending join requests. Some of the public groups turned into private during our collection and we need to send requests to join. Our collection of the groups ranges from November 1, 2017 to August 7, 2019. We collected a total of 47,148 posts created by 6,260 Facebook accounts. Figure 2 shows the number of posts over the collection period, which indicates the overall activity of these review groups over time.

Douban: Sellers create interest groups in Douban to share review exchange information. We collect the posts from ten incentivized review groups in Douban from May 1, 2015 to August 7, 2019. It covers all the groups that we can obtain through the Douban search engine. We collect a total of 3,762 posts from 1,226 authors and obtain 1,031 WeChat accounts in these posts. Figure 3 shows the

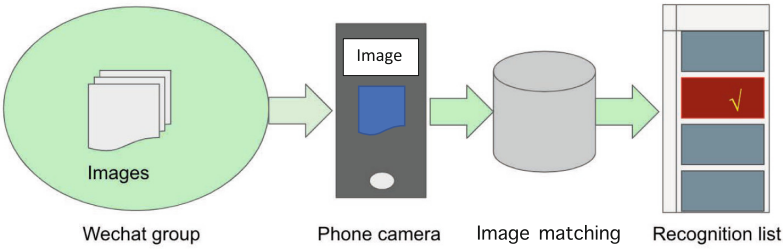


Fig. 6. Amazon product collection. The product images are collected in the WeChat group. Then, we manually search them on Amazon's App via the "camera search" function to identify the corresponding products. If identified in the recognition list, the URLs are collected from the App.

number of posts over time. We find that the incentivized review groups have been becoming increasingly active over time.

WeChat: WeChat group is an ideal place for sellers to broadcast their products since it is private and also offers convenience for further processing and making payments. We send request to join one WeChat group found on Douban and collect the review requests and members' responses over a month from July 7, 2019 to August 7, 2019. In this group, one broker is posting products for several sellers. Figure 4 shows the number of products over time.

3.2 Product Collection

For the purpose of protecting them from detection, sellers are not publishing the URLs of products, but only images and a short introduction. It poses a challenge to the collection of product information involved in the incentivized review groups. To this end, we employ image recognition to collect the corresponding products with the images collected from the group, as shown in Fig. 6. Specifically, we utilize the *Camera Search* feature on Amazon's iOS App to search the products. When searching the products, Amazon may typically pop up a list of products whose descriptions include the image captured from the camera. As such, we need to manually check the product images to ensure that we recognize the correct product in review groups. Note that sellers will copy some parts of product images from other sellers, but scarcely copy all of them. Therefore, we were able to identify such products collected from incentivized review groups. In total, we successfully identify 93 products with image recognition from about 200 products posted in the incentivized review group. We then collect the reviews and product information on these products. For the groups on WeChat, due to significant manual effort and the fact that the review tasks are largely duplicated among the groups, we collect the posts from the largest group which has the most tasks.

Summary. From the above dataset, we find that different platforms play different roles in review groups. The review groups on Facebook are similar to the blackboards, where a set of sellers can post their products directly. In our dataset, there are more than 6,000 sellers who posted products. In the review groups on Douban, most of the posts are to educate sellers on how to obtain incentivized reviews and advertise the brokers who can help sellers. In the review groups on WeChat, there exists a single broker who owns the group. The broker acquires seller members and customer members in many different ways, such as advertisements in Douban. Comparing with review groups on Facebook and Douban, the groups on WeChat are private and hence make members feel sort of close to each other.

4 Measurement

In this section, we examine the collected dataset and characterize incentivized review groups in terms of the group members, review requests, and products.

4.1 Group Members

Figure 5 plots the histogram of member numbers for incentivized review groups we collected from Facebook. We observe that some groups attract a large number of group members (including sellers and reviewers). The largest group has more than 40,000 members. Over a month, there are seven groups that have more than 1,000 new members, indicating that these review groups are remarkably attractive and popular. It also implies that fake reviews from incentivized review groups are still in a considerably large scale.

Sellers: Sellers play a central role in the review groups for posting the review requests that attract members to join the groups. Figure 7 plots the number of sellers for all groups. Note that we label a member as a seller if he/she posts any review request. We can see that there are a number of sellers in most of the review groups, even more than 2,000 sellers in the largest group.

Sellers could join multiple review groups to reach more people and obtain more paid reviews. Figure 8 shows the number of groups that sellers join. We can see that roughly 10% of sellers join more than one group and one of the most aggressive sellers even joins nine review groups at the same time.

4.2 Review Requests

The number of review requests posted in a review group indicates how active the review group is. We observe that some of review groups are notably active during our collection. The most active review group in our dataset has roughly 2,500 review requests every single day.

Figure 9 plots the number of review requests posted by sellers. We can see that some of sellers are notably active, posting more than 100 review requests.

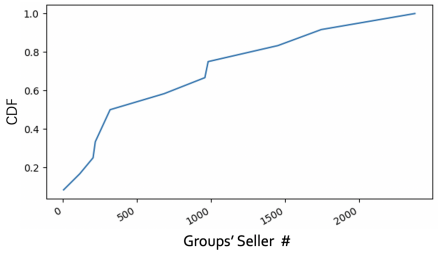


Fig. 7. The number of sellers.

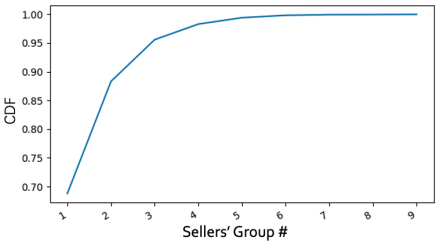


Fig. 8. The number of groups of sellers.

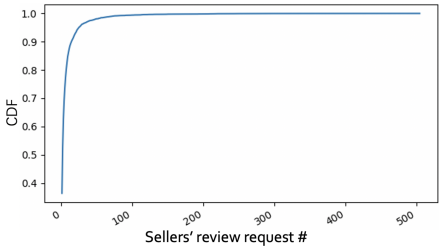


Fig. 9. Review requests of sellers.

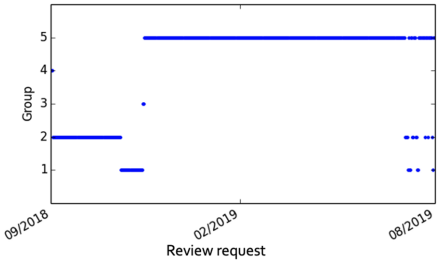


Fig. 10. Review requests across groups.

In Fig. 14, we further depict the number of review requests over time for the seller who posts the most review requests. Interestingly, by labeling the review requests posted in different groups (as shown in Fig. 10), we identify that the seller performed a campaign by focusing on a certain group over a period and then switching to another group later on.

Also, sellers typically send duplicate review requests for some products that are urgent for favorable reviews to boost sales. Figure 11 depicts the number of duplicate requests, which could reach as high as eleven in our dataset, indicating the desperate need for positive reviews.

4.3 Products

Categories. We here investigate the categories of products that stand in need of favorable reviews. *Sports & Fitness* has the most review requests, accounting for 6.88% of total requests. It is followed by *Accessories* and *Computers & Accessories*, making up 5.94% and 5.63%, respectively. We find that 69.5% of the products we collect are labeled as "fulfilled by Amazon", which means that the inventory is kept in Amazon's warehouse and will be packed by Amazon. Not only can the sellers utilize Amazon's facility to facilitate their business, but also the potential benefit is to conceal the place of seller's origin.

Questions and Answers. Customers can ask questions in Amazon and the customers who bought the product may be invited to answer the questions.

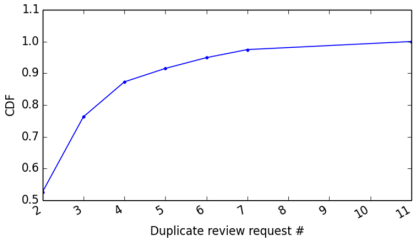


Fig. 11. Duplicate review requests.

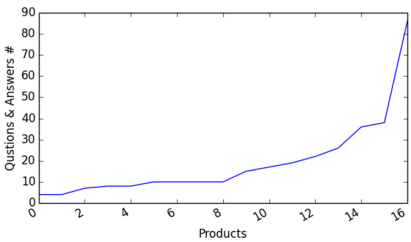


Fig. 12. Questions & Answers.

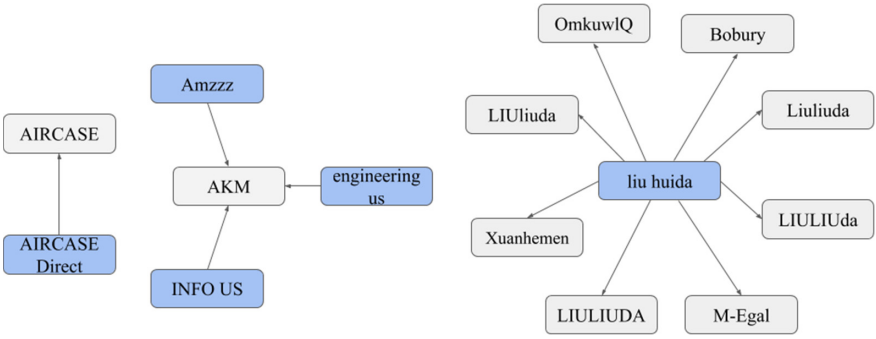


Fig. 13. Relationship types of sellers (blue) and manufacturers (gray). (Color figure online)

Questions & Answers (Q&A) are helpful for addressing customers' concerns and hence could improve the credibility of products. Figure 12 plots the number of the Q&A entries of products collected in the WeChat group. We observe that 16 out of 93 products have at least one Q&A entry, and the largest number of Q&A entries reaches 87. Although the Q&A could also be utilized to promote the products with favorable review requests, we didn't observe the Q&A being manipulated by the review groups.

Sellers and Manufacturers. Manufacturers produce the products which are advertised and listed by sellers on Amazon. We here investigate the relationship between sellers and manufacturers for the products with review requests. Figure 13 shows three different types of relationships. In the left figure, the seller and manufacturer are with a one-to-one relationship and they are usually the same entity. The middle figure reflects a many-to-one model, where multiple sellers work for one manufacturer, and the right figure represents a one-to-many model, where one seller works for multiple manufacturers. Identifying different types of relationship would be useful to understand who is launching the campaign. For example, a many-to-one model implies that it is manufacturers rather than Amazon sellers to request favorable reviews.

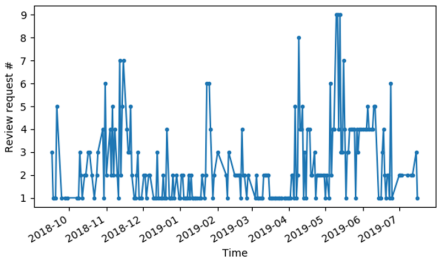


Fig. 14. The number of review requests of the most active seller.

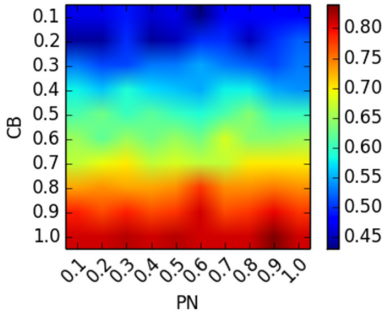


Fig. 15. Varying the weights of CB and PN for LPA method.

4.4 Strategies to Evade Detection

Private Channels. Review groups may operate in private channels, such as the chat groups on WeChat and private groups on Facebook. WeChat group is only visible to group members, and hence it perfectly fits the requirement of the underground business. When we joined the review group on WeChat, it had only about 200 members but quickly reached the maximum limit of 500 members two months later. During the period, most of the new members were invited by the members in the group. The private groups on Facebook are covert and also require permission to join. Due to the effort to discover these groups, the private groups attract the members who are enthused about free product or compensation on the incentivized reviews. Also, the detection of review groups in these private channels is difficult to reach a large scale, and sellers can easily transfer to other review groups.

Without Sharing URLs. Even though sharing URLs of products could simplify the process of review requests and attract more customers, sellers always conceal the URLs of products in the review groups. Even in personal conversations, they are not willing to provide product URLs. The reason is that the URLs from Amazon may include referral information that can be utilized to track the source of sellers. If a number of customers visit a certain product with the same URL that refers to the seller, Amazon can detect the anomaly and probably ban the seller. Concealing URLs in review groups could bring a challenge to our study, which hurdles the collection of products with review requests as well as paid reviews. We utilize an Amazon image recognition procedure to overcome the barrier (Sect. 3).

5 Detecting Incentivized Review Groups with Co-review Graphs

In this section, we model the reviewers as co-review graphs and refer to the identification of incentivized review groups as a community detection problem.

We then employ the graph analysis method to explore the detection. With a "gold standard" dataset collected from real incentivized review groups, we evaluate different community detection algorithms. We also perform a retrospective study on an Amazon review dataset ranging from 1996 to 2018 [10,17].

5.1 Model

We model the reviewers as an undirected graph $G = (V, E)$, where each node $v_i \in V$ represents a reviewer and each edge $\{v_i, v_j\} \in E$ represents a bilateral relationship between v_i and v_j, which indicates that both v_i and v_j write reviews for at least one product. Therefore, we refer to the undirected graph as a *co-review graph*. In the graph, there are $n = |V|$ nodes and $m = |E|$ edges.

To detect the review groups, we employ the graph analysis to detect the communities in the graph and evaluate how accurately the identified communities reflect incentivized review groups. There are more edges inside the communities than the rest of the graph, and the nodes in the same community are considered to be similar to each other. Therefore, the communities of a co-review graph can reveal the cooperation pattern of reviewers in a review graph.

Features. To take various features into our detection, we construct multiple co-review graphs based on different features, such as frequency of co-review and co-review in bursts. Co-review graphs derived from those different features can further improve our detection. Specifically, we consider the following features to construct co-review graphs to perform the community detection:

Frequency of Co-review: The frequency of co-review between two reviewers is one of the most important features for indicating the probability of them belonging to the same incentivized review group. There is no conclusion to draw if two reviewers only occur in one product together. If they occur in more than one product, it is likely that they belong to the same review group, especially when they occur more than three times together. Here, we construct the graph with reviewers occurring more than two times together.

Co-review in Bursts: By checking the time series of reviews of the products that have incentivized reviews, we observe that there exist evident bursts while the products requesting incentivized reviews. Then, we employ Kleinberg's algorithm [12] to detect the burst in the time series. The algorithm models review number in a time series as an infinite hidden Markov model. With the identification of bursts, we collect the co-review of reviewers in the bursts. For the reviewers of review groups, they are required to post the most favorable reviews to obtain free products or compensation. Therefore, we also check the rating of reviewers in the bursts, *e.g.*, five stars in Amazon.

Posting Nearness in Time: The closer in time two reviewers post their reviews, the more possible they belong to the same review group. Also, there exist some legitimate reviews that occur very close to each other. Based on the purchase confirmations in the group, we can identify that most purchases by

incentivized reviewers were made within two days, given these products are all Amazon Prime products with free 2-day delivery. By collecting reviewers posting positive reviews (five stars) in the same product within two days, we construct the co-review graph in terms of posting nearness in time.

Composing Multiple Graphs. We then denote the graph from the frequency of co-review as **FC graph**, the graph from co-review in bursts as **CB graph**, and the graph from posting nearness in time as **PN graph**. Multiple graphs derived from different features are complementary to each other. For example, CB graphs have some important edges although two nodes of these edges co-occur only once and hence they do not exist in the FC graph. As such, we compose multiple graphs according to the following equation:

$$\mathcal{G} = \mathcal{G}_{FC} + W_{CB}\mathcal{G}_{CB} + W_{PN}\mathcal{G}_{PN}. \tag{1}$$

First, we derive the FC graph by taking into account all pairs of nodes co-occurring more than once. Then, we compose the CB graph into the FC graph by adding edges that have at least one node in the FC graph with weight W_{CB}, which measures the importance of co-review in burst feature. Similarly, we compose the PN graph into the FC graph with weight W_{PN}, which denotes the importance of posting nearness in time feature.

5.2 Community Detection

Dataset. For further exploring the community of incentivized review groups, we collect the products posted in the review groups, including seller information, all reviews, and questions & answers from customers. As mentioned in Sect. 3, sellers always conceal the products' URLs and are not willing to provide them even in personal conversation. We utilize an image recognition procedure to identify the products on Amazon. We identify 93 products posted in review groups by searching product images of more than 200 products. These identified products belong to 48 sellers. We further collect 531 products from these sellers. We find that sellers usually cooperate with more than one incentivized review groups and select different products for different time periods or different incentivized review groups. Therefore, some products from them are likely to be posted in the incentivized review groups that we do not have access or the periods out of our collection.

"Gold Standard" Dataset: Since we have knowledge of products and reviews posted by the incentivized review groups, we can construct a "gold standard" dataset with these factual incentivized reviews as ground-truth. The dataset consists of 764 incentivized reviews from 737 reviewers. With the dataset, we extract the co-review connections of reviewers and evaluate the community detection algorithms. As a result, we obtain 5,950 co-review connections from the "gold standard" dataset.

Table 1. AMI among algorithms.

	CPM	Louvain	LPA	Infomap
CPM	–	0.80	0.79	0.14
Louvain	–	-	0.83	0.12
LPA	–	–	–	0.14
Infomap	–	–	–	–

Methods. Next, We explore four different community detection methods to detect incentivized review groups:

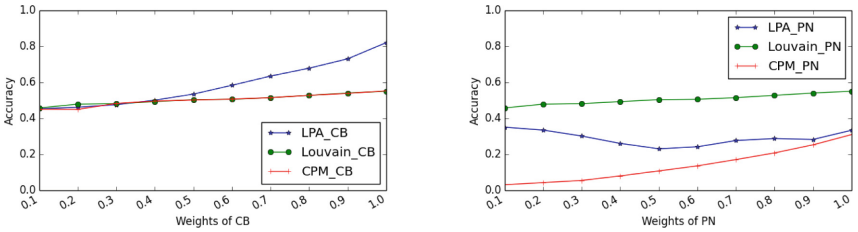
Clique Percolation Method (CPM): The clique percolation method [20] constructs the communities from k-cliques, which correspond to fully connected sub-graphs of k nodes. Two k-cliques are considered adjacent if they share $(k-1)$ nodes and a union of adjacent k-cliques form a community.

Louvain: Louvain method [4] first finds small communities by optimizing modularity locally on all nodes and then group small communities into nodes. It repeats above the two steps until achieving the optimal modularity. Modularity is a scale value between -1 and 1 that measures the density of edges inside communities to edges outside communities. Optimizing this value theoretically results in the best possible grouping of the nodes in a given network.

Label Propagation Algorithm (LPA): The label propagation algorithm [21] works by propagating labels throughout the network to form communities, where a small subset of nodes have been pre-assigned with labels. Intuitively, a single label can quickly become dominant in a densely connected group of nodes, but it is difficult to cross a sparsely connected region. The nodes that end up with the same label can be considered to be in the same community.

Infomap: Infomap [23] uses community partitions as a Huffman code that compresses the information about a random walker exploring a graph. A random walker exploring the network with the probability that the walker transits between two nodes given by its Markov transition matrix. Once the random walker enters the densely connected regions of the graph, it tends to stay there for a long time, and movements between the regions are relatively rare, which allows us to generate Huffman codes with modularity information. A modular description of a graph can be viewed as a compression of the graph topology.

Clustering Comparison. Here, we compare the results by employing above community detection algorithms to identify the communities corresponding to incentivized review groups. First, we compare the results from different algorithms by measuring the Adjusted Mutual Information (AMI) among different algorithms. AMI accounts for how similar two community detection results are to each other. As shown in Table 1, we can see that the detection results of those algorithms are similar to each other, especially the LPA and Louvain method.



(a) Varying CB weights. (b) Varying PN weights.

Fig. 16. Varying different weights.

However, the result of Infomap algorithm is remarkably distinct from other algorithms. After careful inspection, we find that Infomap groups most of nodes to one huge community. Therefore, we consider that Infomap is not suitable for this problem. Note that we empirically set $k = 4$ in CPM for the results in Table 1. For example, for $k = 3$, we observe that the AMI between CPM and Louvain drops to 0.43 and the AMI between CPM and LPA falls to 0.40. This inconsistency indicates that $k = 3$ may underperform comparing with $k = 4$.

Varying Weights of Composing Graphs. We then utilize the "gold standard" dataset to evaluate the accuracy of the above algorithms. The accuracy is measured by the proportion of the factual connections extracted from the "gold standard" dataset that are correctly identified by a community detection algorithm. When considering one type of graph alone, the accuracy is prohibitively low. For example, the FC graph produces the accuracy of 0.46 and 0.35 for Louvain and LPA method.

To improve the accuracy, we examine the impact by composing the PN and CB graphs into the FC graph, respectively. First, we empirically determine the composing weights in Eq. (1) by measuring the importance of the PN graph and CB graph, as illustrated in Fig. 16. We can see that composing the CB graph can significantly improve the accuracy of LPA method. When fully composing the CB graph, LPA's accuracy achieves 81%. Meanwhile, the CPM and Louvain method only gains trivial improvement with the CB graph. On the other hand, when composing the PN graph into FC graph, the CPM method gains constant improvement but remains lower than other methods, while the accuracy of LPA method first drops and then rises up. Overall, composing the CB graph achieves higher accuracy than the PN graph. It is probably because although the PN graph is roughly 10 times bigger than the FC graph and CB graph, it may also carry a bunch of unwanted nodes and edges. In the end, we here choose the LPA method to further conduct community detection.

Furthermore, we vary the weights of the CB graph and PN graph at the same time to explore an optimal weight combination of LPA method to achieve the best performance. Figure 15 depicts the heat map of accuracy, where the sidebar represents the scale of accuracy. We can see that ($W_{CB} = 1.0$, $W_{PN} = 0.9$)

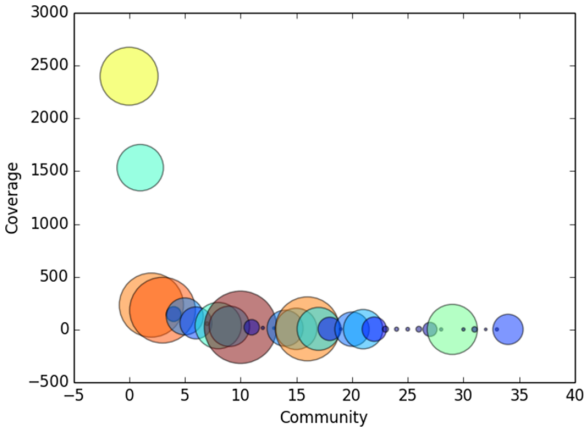


Fig. 17. An example of communities. The size of a circle specifies the number of nodes in the community, and the coverage of communities (y-axis) indicates how many edges from our "gold-standard" dataset they cover.

achieves the best accuracy of 85% in our experiment, although it is just a bit higher than ($W_{CB} = 1.0, W_{PN} = 0.1$). It confirms that the CB graph remarkably improves the community partition comparing with the PN graph.

Communities as Incentivized Review Groups. We then investigate the distribution of communities, which could also be used to reflect the performance of community detection method. For example, if the biggest community identified by the community detection method includes most of nodes and covers nearly all of edges from our "gold-standard" dataset, this detection method would achieve a high accuracy but essentially useless. Therefore, we prefer a balanced community detection method. We select LPA method that achieves the best performance above as an example and plot the distribution of communities in Fig. 17. The area of the circle indicates the number of nodes in the community and the coverage of communities (y-axis) represents how many edges from our "gold-standard" dataset they included. We can see that LPA method partitions notably balanced communities. The left two communities which cover most of edges from our "gold-standard" dataset are apparently the communities engaged in incentivized review groups. They are both of moderate size, 1015 and 654 respectively. We then further inspect the nodes of these two communities in the following Sect. 5.3.

5.3 Reviewer Profiles

For the reviewers in two communities mentioned above, we collect their public profiles from Amazon and investigate their ranking, the number of reviews, and the number of helpful votes. Amazon ranks reviewers by a private algorithm,

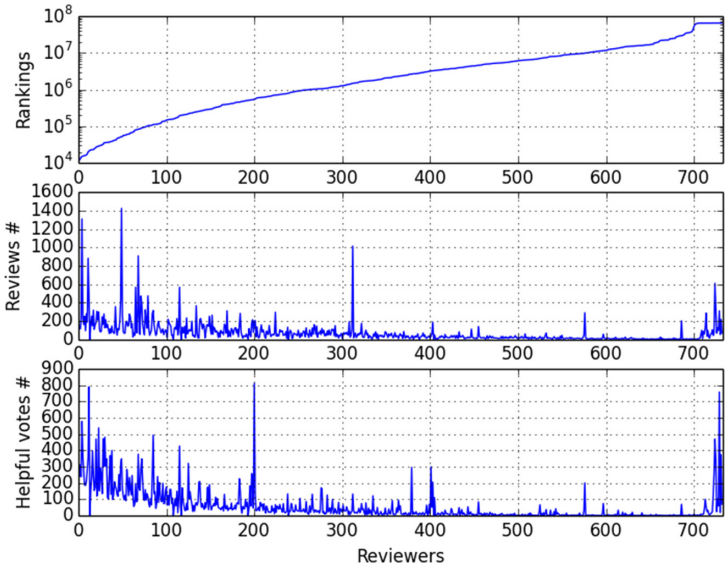


Fig. 18. Ranking, review #, and helpful vote # of reviewers.

where a smaller ranking score represents a higher reputation. Reviewers with higher reputation would be highly preferred by sellers who ask for incentivized reviews, since their reviews would be more authentic and trustworthy. The number of reviews written by a reviewer demonstrates how active the reviewer is, while the number of helpful votes a reviewer received can reflect how helpful the reviews of the reviewer are. In other words, it suggests to what extent the reviewer helps other customers. Figure 18 depicts ranking (top), number of reviews (middle), and number of helpful votes (bottom) of reviewers.

We can observe that, in the left part of those figures, reviewers with higher reputation also produce more reviews and receive more helpful votes. In the middle part, some spikes represent that a few reviewers also have an outstanding amount of reviews or helpful votes. In the right part, some reviewers with relatively lower reputation have an extraordinary amount of both reviews and helpful votes. After inspecting these reviewers, we find that they post a number of reviews within a short period and some of these reviewers actually obtain helpful votes from other customers reciprocally or from suspicious accounts.

5.4 A Retrospect of Amazon Dataset

We here conduct a retrospective study of Amazon review groups with the public datasets [10,17]. The dataset [10] contains product reviews and metadata from Amazon, including 142.8 million reviews from May 1996 to July 2014 (we refer to it as 2014 dataset). We construct the co-review graph and find that there are only 1,022 reviewers in the co-review graph. It indicates that incentivized

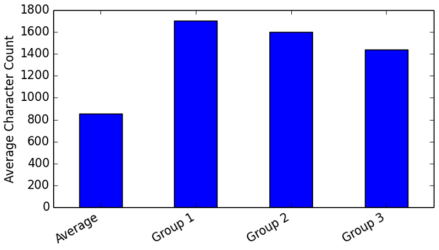


Fig. 19. Average character count per review.

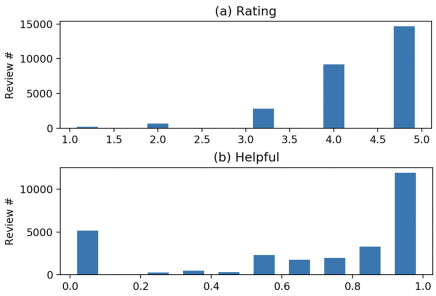


Fig. 20. Rating and helpful index.

review groups were not on an extensive scale before 2014. Then, with the updated version of dataset [17] (we refer to it as 2018 dataset), we extract the new reviews ranging from 2015 to 2018 and construct the co-review graph, which includes 90.3 million reviews. It turns out that we obtain a co-review graph with 197,087 reviewers, which is significantly higher than the 2014 dataset.

Next, we apply the LPA community detection method for processing the co-review graph with the frequency of co-reviews. We identify 31 groups in the 2014 dataset and 6,278 groups in the 2018 dataset. To further investigate the groups, we inspect three largest groups, which contain 115, 109, and 71 nodes, respectively, labeled as "Group1", "Group2", and "Group3". Figure 19 plots the average character count per review across different groups. The left bar, "Average" represents the average character count over all the reviews in the dataset. We can see that these three groups have remarkably more characters than the average, which implies that the reviewers from these groups are possibly professional critics who are invited to write professional reviews.

Figure 20 shows the distribution of rating and helpful index of the largest group, *i.e.*, Group1. We observe that there exist a number of average reviews less than 4 and also a spate of reviews' helpful index less than 0.5, which implies that the reviews are not considerably biased. We also inspect the review timestamps and find no anomaly.

Summary: By comparing the datasets ranging from 1996 to 2014 with the dataset ranging from 2015 to 2018, we can see that the co-review graph has shown a significant shift since 2015, indicating that the incentivized review group has become a serious issue for online marketing platforms such as Amazon.

6 Related Work

6.1 Spam Review Detection

Yao *et al.* [29] presented a potential attack against online review systems by employing deep learning to automatically generate fake reviews. They also proposed countermeasures against these fake reviews. Wang *et al.* [24] built review

graphs to capture the relationships among reviewers, reviews, and stores, and then quantified the trustiness of reviewers. Zheng *et al.* [30] attempted to detect elite Sybil fake reviews in Sybil campaigns. Rayana *et al.* [22] exploited behavioral data, text data, and relational data to detect spam reviews and reviewers. Ott *et al.* [18,19] detected deceptive reviews from both positive and negative sentiment review datasets. Song *et al.* [7] investigated syntactic stylometry for deception detection. Li *et al.* [13] detected deceptive opinion spam across different domains. Mukherjee *et al.* [16] examined filtered reviews of Yelp and inferred their filtering algorithms. Fusilier *et al.* [8] employed character n-gram features to detect deceptive opinion spam. Harris *et al.* [9] examined a variety of human-based, machine-based, and hybrid assessment methods to detect deceptive opinion spam in product reviews. In [11], Jamshidi *et al.* examined the explicitly incentivized reviews which state their incentives explicitly in the reviews. Different from [11], we investigate the underground economy of incentivized reviews across different social networking platforms and propose a detection method for the incentivized review groups. Also, Mukherjee *et al.* [15] identified opinion spam groups based on a set of spam behavior indicators. These spam behavior indicators could also be applicable to improve our detection of incentivized review groups. Xie *et al.* [25] utilized temporal patterns to detect singleton review spam by identifying the time windows when spam reviews are likely to happen. However, such a method is not suitable for detecting the incentivized review groups since spammers actually collude in a collection of different products. As such, we propose a detection method based on the co-review graph, which can correlate these spammers from different products.

6.2 Reputation Manipulation

In online markets, sellers' reputation is closely related to profitability. Dishonest sellers have been reported to maneuver the reputation system by manipulating the transaction history. Xu *et al.* [28] investigated the underground market by which sellers could easily harness human labor to conduct fake transactions for improving their stores' reputation. They referred to this underground market as Seller-Reputation-Escalation (SRE) markets. Cai *et al.* [5] employed reinforcement learning methods to detect reputation manipulation in online markets. Li *et al.* [14] investigated the manipulation of mobile app reputation by leveraging crowdsourcing platforms. In [6], the authors exploited the unusual ranking change patterns of apps to identify promoted apps and detected the collusive groups who posted high app ratings or inflated apps' downloads. In addition, Xie *et al.* [26,27] inspected the underground market where mobile app developers could misuse positive reviews illegally or manipulate the rating collusively. They also analyzed the promotion incentives and characteristics of promoted apps and suspicious reviews. By contrast, our work focuses on the manipulation of reputation in online markets, which leverages incentivized review groups. The existing detection methods are unable to address this emerging manipulation problem. Therefore, we propose a novel detection method based on the co-review graph for effective defense.

7 Conclusion

In this paper, we revealed a new online reputation manipulation problem existed in the incentivized review groups on Amazon. We first investigated incentivized review groups across different platforms to understand the breadth of the problem and conducted a measurement study by considering group members, review requests, and products. After the measurement study, we proposed a detection method based on co-review graphs. We leveraged the community detection methods to locate the suspicious communities from the co-review graphs with high accuracy. While evaluating our detection method, we also constructed a "gold standard" incentivized review group dataset, which provides the critical ground truth for further study on incentivized reviews.

Acknowledgment. We would like to thank our shepherd Mohammad Mannan and the anonymous reviewers for their detailed and insightful comments, which help to improve the quality of this paper. This work was supported in part by the U.S. ARO grant W911NF-19-1-0049 and NSF grant DGE-1821744.

References

1. Amazon bans incentivized reviews (2016). https://www.usatoday.com/story/tech/news/2016/10/03/amazon-bans-incentivized-reviews/91488702/
2. Amazon consumer review policy (2016). https://www.amazon.com/gp/help/customer/display.html?nodeId=201967050
3. Consumer review survey (2017). https://www.brightlocal.com/research/local-consumer-review-survey/
4. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. Theo. Exp. **2008**(10), P10008 (2008)
5. Cai, Q., Filos-Ratsikas, A., Tang, P., Zhang, Y.: Reinforcement mechanism design for fraudulent behaviour in e-commerce. In: The 32nd AAAI Conference on Artificial Intelligence (2018)
6. Chen, H., He, D., Zhu, S., Yang, J.: Toward detecting collusive ranking manipulation attackers in mobile app markets. In: The 2017 ACM on Asia Conference on Computer and Communications Security (2017)
7. Feng, S., Banerjee, R., Choi, Y.: Syntactic stylometry for deception detection. In: The 50th Annual Meeting of the Association for Computational Linguistics: Short Papers, vol. 2, pp. 171–175. Association for Computational Linguistics (2012)
8. Fusilier, D.H., Montes-y-Gómez, M., Rosso, P., Cabrera, R.G.: Detection of opinion spam with character n-grams. In: Gelbukh, A. (ed.) CICLing 2015. LNCS, vol. 9042, pp. 285–294. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-18117-2_21
9. Harris, C.G.: Detecting deceptive opinion spam using human computation. In: Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence (2012)
10. He, R., McAuley, J.: Ups and downs: modeling the visual evolution of fashion trends with one-class collaborative filtering. In: Proceedings of the 25th International Conference on World Wide Web (WWW), pp. 507–517 (2016)
11. Jamshidi, S., Rejaie, R., Li, J.: Trojan horses in amazon's castle: understanding the incentivized online reviews. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (2018)

12. Kleinberg, J.: Bursty and hierarchical structure in streams. Data Min. Knowl. Disc. **7**(4), 373–397 (2003)
13. Li, J., Ott, M., Cardie, C., Hovy, E.: Towards a general rule for identifying deceptive opinion spam. In: The 52nd Annual Meeting of the Association for Computational Linguistics, vol. 1 (2014)
14. Li, S., Caverlee, J., Niu, W., Kaghazgaran, P.: Crowdsourced app review manipulation. In: The 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (2017)
15. Mukherjee, A., Liu, B., Glance, N.: Spotting fake reviewer groups in consumer reviews. In: The 21st International Conference on World Wide Web (WWW) (2012)
16. Mukherjee, A., Venkataraman, V., Liu, B., Glance, N.S.: What yelp fake review filter might be doing? In: The International AAAI Conference on Web and Social Media (ICWSM) (2013)
17. Ni, J., Li, J., McAuley, J.: Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In: The Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (2019)
18. Ott, M., Cardie, C., Hancock, J.T.: Negative deceptive opinion spam. In: The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2013)
19. Ott, M., Choi, Y., Cardie, C., Hancock, J.T.: Finding deceptive opinion spam by any stretch of the imagination. In: The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1. Association for Computational Linguistics (2011)
20. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature **435**(7043), 814 (2005)
21. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E **76**(3), 036106 (2007)
22. Rayana, S., Akoglu, L.: Collective opinion spam detection: bridging review networks and metadata. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2015)
23. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proc. Nat. Acad. Sci. **105**(4), 1118–1123 (2008)
24. Wang, G., Xie, S., Liu, B., Philip, S.Y.: Review graph based online store review spammer detection. In: 2011 IEEE 11th International Conference on Data Mining (2011)
25. Xie, S., Wang, G., Lin, S., Yu, P.S.: Review spam detection via temporal pattern discovery. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2012)
26. Xie, Z., Zhu, S.: AppWatcher: unveiling the underground market of trading mobile app reviews. In: The 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks, p. 10 (2015)
27. Xie, Z., Zhu, S., Li, Q., Wang, W.: You can promote, but you can't hide: large-scale abused app detection in mobile app stores. In: The 32nd Annual Conference on Computer Security Applications (2016)
28. Xu, H., Liu, D., Wang, H., Stavrou, A.: E-commerce reputation manipulation: the emergence of reputation-escalation-as-a-service. In: International Conference on World Wide Web (WWW), pp. 1296–1306 (2015)

29. Yao, Y., Viswanath, B., Cryan, J., Zheng, H., Zhao, B.Y.: Automated crowdturfing attacks and defenses in online review systems. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (2017)
30. Zheng, H., et al.: Smoke screener or straight shooter: detecting elite Sybil attacks in user-review social networks. In: NDSS (2018)